Real Analysis

P. Ouwehand

Department of Mathematical Sciences
Stellenbosch University

Contents

Chapter 1

An Axiomatic Development of the Real Number System

Traditionally, a rigorous course in Analysis progresses (more or less) in the following order:

$$\begin{array}{lcl} \textit{sets,} & & \textit{limits,} \\ \textit{mappings,} & \Rightarrow & \textit{continuous} \\ & & \textit{functions} \end{array} \quad \Rightarrow \quad \textit{derivatives} \quad \Rightarrow \quad \textit{integration}$$

On the other hand, the historical development of these subjects occurred in reverse order:

$$\begin{array}{lcl} \textit{Cantor 1875,} & & \textit{Cauchy 1821,} \\ \textit{Dedekind} & \Leftarrow & \textit{Weierstrass} \end{array} \quad \Leftarrow \quad \begin{array}{l} \textit{Newton 1665} \\ \textit{Leibniz 1675} \end{array} \quad \Leftarrow \quad \begin{array}{l} \textit{Archimedes} \\ \textit{Kepler1615} \\ \textit{Fermat 1638} \end{array}$$

— From *Analysis by Its History*, by E. Hairer and G. Wanner, Springer 1996

1.1 Why we need axioms

Consider the following questions:

Question 1: Many years ago, you were taught the following algorithm for multiplying two numbers:

$$\begin{array}{r} 23 \\ \underline{\times\ 17} \\ 161 \\ \underline{230\ \ } \\ 391 \end{array}$$

Why does this algorithm work?

Question 2: Why is $-1 \times -1 = 1$? Alternatively, why is the product of two negative numbers a positive number?

If you think that these are silly questions, think again. The answers to these questions are *not* obvious. You are merely so used to the answers that the questions never occur to you.

An explanation for why the multiplication algorithm works might go along the following lines:

$$\begin{aligned} 23 \times 17 &= 23 \cdot (7 + 10) \\ &= 23 \cdot 7 + 23 \cdot 10 \\ &= (20 + 3) \cdot 7 + (20 + 3) \cdot 10 \\ &= [20 \cdot 7 + 3 \cdot 7] + [20 \cdot 10 + 3 \cdot 10] \\ &= [140 + 21] + [200 + 30] \qquad\qquad = 161 + 230 \\ &= 391 \end{aligned}$$

To do this calculation, we performed the following operations:

(i) We used the fact that $a \cdot (b + c) = a \cdot b + a \cdot c$ several times.

(ii) We retrieve certain results, like $3 \cdot 7 = 21$, from memory. Such results were learnt by rote, in the form of multiplication tables. Thus all the values of $a \times b$ for $1 \leq a, b \leq 10$ are stored in a mental look–up table.
The values in the look–up table were determined *empirically*, i.e. by observation. To see that $7 \times 8 = 56$, take 8 small bags, each containing 7 stones, and empty them into a big bag. If you now count the number of stones in the big bag, you will get 56. That's just a fact that's been observed over and over again, in many different places and at many different times.

(iii) We use the fact that multiplying a number by 10 is accomplished by adding a zero to the end of that number. Thus $20 \cdot 10 = 200$.

(iv) To calculate the value of a term such as $20 \cdot 7$ (which is not in the mental look–up table), we have to argue that $20 \cdot 7 = 7 \cdot 20 = 7 \cdot (2 \cdot 10) = (7 \cdot 2) \cdot 10 = 14 \cdot 10 = 140$. Thus, in addition to the look–up table and the multiply–by–ten rule, we also used the following facts about multiplication: $a \cdot b = b \cdot a$, and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

(v) We used another algorithm (also learnt long ago) for adding numbers, such as $161 + 230$. Try and justify that algorithm yourself.

As you can see, in order to explain why the multiplication algorithm works, you need to invoke quite a few simpler results about addition and multiplication. Question 1 is not as obvious as it looks! As for Question 2, you should be able to explain why $-1 \times -1 = 1$ by the end of this chapter.

Now note the following (empirically verifiable) facts: Human beings have a certain intuition (or idea) about non–physical objects called numbers. These numbers can be combined in various ways to form new numbers, e.g. they can be added and multiplied. Moreover, there are some simple rules which govern the combination of numbers, e.g.

(i) The product of two numbers does not depend on where or when the multiplication is performed.

(ii) $a + b = b + a \qquad ab = ba$

(iii) $a + (b + c) = (a + b) + c \qquad a(bc) = (ab)c$

(iv) $a(b+c) = ab + ac$

et cetera. Our aim is now to find a set of rules, or *axioms*, which completely captures our intuition about the arithmetic of the reals. In other words, we seek a set of rules such that

(1) The rules are in accord with our intuition about arithmetic.

(2) The set of rules is sufficiently rich that any informal, intuitive arithmetic argument can be made *formal*: we can reach the same conclusion by applying no intuition at all, but just the formal rules (axioms).

Why do we need axioms? For several reasons.

- Axioms tend to be simple, and most people will accept them as in agreement with their intuition. Thus the axioms are a common starting point for all people. People who disagree about the axioms are probably talking about different things.

- The agreed–upon rules can be applied over and over again, to arbitrary levels of complexity. Any two people who agree on the (simple) axioms *must* also agree on the (complicated) conclusions that may be reached by formal application of those axioms.

  On the other hand, intuition becomes less and less reliable as we increase the level of complexity, and thus conclusions obtained solely by a intuition are more suspect.
  For example, you and I may agree that Euclid's 5 axioms for geometry are in accordance with our intuition of *space*. These axioms are simple, and difficult to disbelieve. *You* may have a powerful intuition, however: You intuit that the square of the (length of) the hypotenuse of a right–angled triangle is equal to the sum of the squares of the other two sides. But *my* intuition is far less developed than yours: I just don't see it, and so I don't believe you. Should you provide a step–by–step argument, starting from our common ground (the 5 axioms), using only commonly agreed rules, I will be forced to admit that your intuition is correct. In this way, I can *verify* the truth of your assertion myself, and don't just have to take your word for it.

- If we use the axiomatic method, we are constantly aware of our assumptions. It therefore becomes much simpler to discern similarities and differences between various mathematical objects and operations. This will make the arguments *portable* (in the Computer Science sense — arguments (computer code) can easily be moved from one problem to (platform) to another).

- Finally, axioms allow us to *circumvent metaphysical speculation* about the nature and existence of mathematical objects. What, for example *is* a real number? Is it an irreducible, or is it made up of simpler things? This question was first given a satisfactory answer in 1872. Indeed, it was given *two* different but satisfactory answers in that year, by Dedekind and Cantor. In each case, the real numbers are "constructed" from some previously constructed, simpler, objects, e.g. the rational numbers.

  Thus there is no single answer to the question: "What is a real number?". But the exact *nature* of the reals is unimportant for mathematical purposes. What is important is how they *behave*, i.e. how they can be recombined, using various operations, to form new numbers. The axioms are essentially just a description of such behaviour, and though the three constructions disagree about the essential nature of the reals, they *do* agree on how they behave.

Example 1.1.1 Series are not the same as sums In high school you learnt that

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \cdots = 2$$

A typical argument to prove this fact, which you probably used a number of times in your first year calculus course, might go as follows:

Let $S := 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots$. Then

$$\begin{aligned} S = 2S - S &= 2(1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{16} \ldots) - (1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \ldots) \\ &= (2 + 1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \ldots) - (1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \ldots) \\ &= 2 + [(1-1) + (\tfrac{1}{2} - \tfrac{1}{2}) + (\tfrac{1}{4} - \tfrac{1}{4}) + (\tfrac{1}{8} - \tfrac{1}{8}) + \ldots] \\ &= 2 \end{aligned}$$

Consider now the series

$$A := 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \tfrac{1}{5} - \tfrac{1}{6} + \ldots$$

It can be shown — and we will do this later — that the above series converges to $A := \ln 2$. If we rearrange the terms according to the recipe "One odd, then two evens" we obtain

$$\begin{aligned} A &= (\underbrace{1 - \tfrac{1}{2}}_{\frac{1}{2}} - \tfrac{1}{4}) + (\underbrace{\tfrac{1}{3} - \tfrac{1}{6}}_{\frac{1}{6}} - \tfrac{1}{8}) + (\underbrace{\tfrac{1}{5} - \tfrac{1}{10}}_{\frac{1}{10}} - \tfrac{1}{12}) + \ldots \\ &= (\tfrac{1}{2} - \tfrac{1}{4}) + (\tfrac{1}{6} - \tfrac{1}{8}) + (\tfrac{1}{10} - \tfrac{1}{12}) + \ldots \\ &= \tfrac{1}{2}(1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \tfrac{1}{5} - \tfrac{1}{6} + \ldots) \\ &= \tfrac{1}{2}A \end{aligned}$$

i.e. $\ln 2 = \frac{1}{2} \ln 2$ — a contradiction. This example shows that a basic intuition fails: one cannot simply interchange and rearrange the terms in an infinite series and expect to get the same answer, though this is perfectly acceptable for finite sums.
At this point, you should be feeling rather uncomfortable about the series $S := 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots$ — the argument that $S = 2$ also seems to use some rearrangements. We need to examine the notion of *convergence of an infinite series* rather more closely.

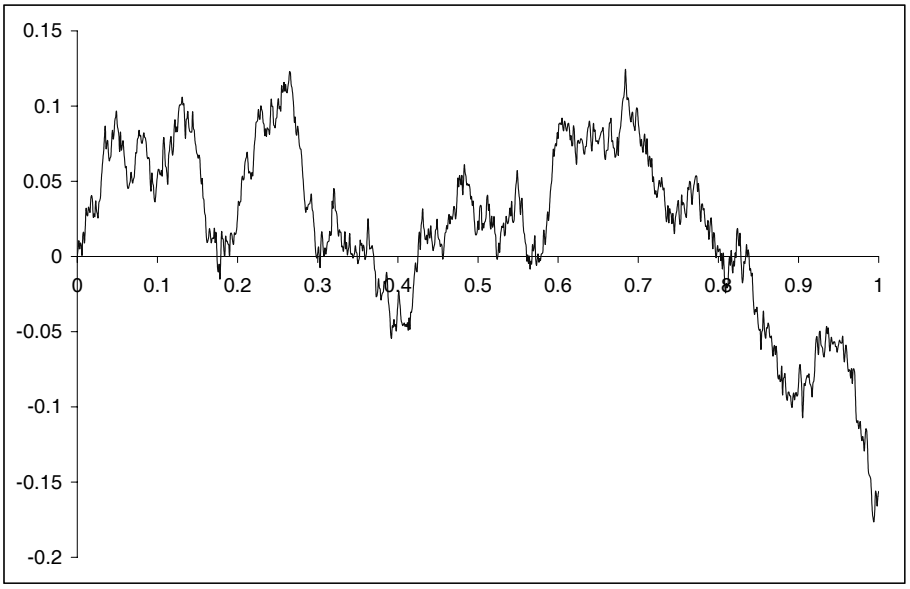
Example 1.1.2 A Useful Monster: Brownian Motion
We consider here an example of a counter–intuitive "monster" which is used extensively in engineering and control theory, as well as economics and finance, to describe random phenomena.
Although a precise definition requires some knowledge of elementary statistics, a *Brownian motion* (also called *Wiener process*) is a random function $B : [0, \infty) \to \mathbb{R}$ defined, roughly, as follows:

(i) $B(t)$ is a continuous function with $B(0) = 0$. The variable t is thought of as *time*.

(ii) Each change $B(t + s) - B(s)$ is a normally distributed random variable, with mean zero and variance equal to the length of the elpapsed time $t + s - s = t$.

(iii) Changes on non-overlapping time intervals are independent of each other.

A graph of a path of a Brownian motion is shown below:



Using some rather difficult analysis, one can then prove that Brownian motion has the following properties:

- Though everywhere continuous, each path is *nowhere differentiable*, i.e. $\frac{dB}{dt} := \lim_{s \to 0} \frac{B(t+s)-B(t)}{s}$ does not exist for any $t \geq 0$. This means that it has no "smooth" bits anywhere.
- The length each path over any finite non–zero interval is *infinite*. In particular, the path shown in the graph has infinite length over the interval $[0, 1]$ (and also over the interval $[0, 10^{-100000000}]$).

Until the mid–nineteenth century, it was widely believed that any continuous function must be differentiable at least somewhere, but Brownian motion provides a counterexample. Though it was used by Louis Bachelier in 1900 to determine the prices of stock options, and by Albert Einstein in 1905 as evidence for the existence of atoms, the mathematical object "Brownian motion" was shown to exist only in 1923, by Norbert Wiener. For such monsters, intuition is useless, and only careful analysis can provide insight.

1.2 Arithmetic of Fields

The aim of this chapter is to present an *axiomatization* of the real number system. We shall do this in three stages:

(1) First we shall discuss the purely arithmetic properties of the reals. The reals form an algebraic system called a *field*. Intuitively, a field is a set in which addition, subtraction, multiplication and division are possible, and obey the usual rules.

(2) Next, we shall discuss the properties of a field equipped with an ordering relation $\leq$.

(3) Finally, we shall add an axiom, the *Completeness Axiom*, which ensures that it is possible to take limits.

The aim is to write down a set of axioms which completely characterize the system of real numbers.

Remarks 1.2.1 Throughout this section, I refer to the reals as though you already know what what they are, and how they behave (which, of course, to a large extent, you do). The more philosophically minded may therefore come to believe that much of the following discussion is *circular*: I *define* the properties of the reals by *observing* the properties of the reals.
That is not the case. We will operate on two levels, the *intuitive* and the *formal*: We have an intuitive idea of real numbers, to which I make frequent appeal. For example, we may think of real numbers as points on a line, or as objects that have a representation of the form $3.14159265\ldots$ (i.e. a decimal representation), to name just two intuitive ideas of number. We then extract the basic properties (axioms) of the real number system from our intuitive notions, leading from the informal level to formal. All theorems, propositions, etc. will be proved from these formal properties. Thus while we *do* use intuitive notions to write down the basic axioms, we *do not* use intuitive notions to prove theorems. For that, we use only the basic axioms.

□

Definition 1.2.2 A *field* is a tuple $\langle F, +, \cdot, -, ^{-1}, 0, 1\rangle$, satisfying all the properties below:

- F is a *set*;
- $+, \cdot$ are *binary operations* on F.

  This means that $+, \cdot$ are functions of two variables on F:

  $$+, \cdot : F \times F \to F$$

  It is customary to write $a + b$ instead of $+(a, b)$, and $a \cdot b$ or ab instead of $\cdot(a, b)$.

- $0, 1$ are *distinct* designated members of F.

  Such elements are also called *nullary operations* on F, or constants. We call 0 the *zero element*, or *additive identity*. Similarly, we call 1 the *unit element*, or the *multiplicative identity*.

- $-, ^{-1}$ are unary operations on F.

  Thus $-, ^{-1}$ are functions from F to F. However, $^{-1}$ is a *partial function*: a^{-1} is defined if and only if $a \neq 0$.

In addition, the operations are required to satisfy the following basic properties (axioms):

(C^+)	$a+b=b+a$	(Commutativity of addition)
(A^+)	$(a+b)+c=a+(b+c)$	(Associativity of addition)
(Id^+)	$a+0=a=0+a$	(Additive identity)
(Inv^+)	$a+(-a)=0=(-a)+a$	(Additive inverse)
($C^\cdot$)	$a\cdot b=b\cdot a$	(Commutativity of multiplication)
($A^\cdot$)	$(a\cdot b)\cdot c=a\cdot(b\cdot c)$	(Associativity of multiplication)
($Id^\cdot$)	$a\cdot 1=a=1\cdot a$	(Multiplicative identity)
($Inv^\cdot$)	$a\cdot(a^{-1})=1=(a^{-1})\cdot a$ when $a\neq 0$	(Multiplicative inverse)
(D)	$a\cdot(b+c)=a\cdot b+a\cdot c$	(Distributivity of $\cdot$ over $+$)

Remarks 1.2.3 We introduce here some basic (and familiar) rules designed to simplify notation: Firstly, we will assume that the operations satisfy the usual order of precedence: $^{-1}$ before $\cdot$ before $-$ before $+$. Thus $ab^{-1}=a\cdot(b^{-1})$ (and not $(ab)^{-1}$), $ab+c=(ab)+c$ (and not $a(b+c)$), etc.
Next, we define the expression "$b-a$" to mean "$b+(-a)$", and we define "a/b" to mean "ab^{-1}".
By (A^+), $(a+b)+c=a+(b+c)$. We may therefore omit the brackets and denote this common value by $a+b+c$. Similarly, "abc" is defined to be the common value of $(ab)c$ and $a(bc)$.
We will also write "a^2b" instead of "aab", "a^{-2}" instead of "$(a^{-1})^2$", etc.

□

Are the above axioms sufficient to characterize the system of real numbers? No. As is demonstrated by the next example, there are many different examples of fields, and each field satisfies the axioms (C^+),(A^+),... (D).

Examples 1.2.4 (1.) The set $\mathbb{Q}$ of rational numbers with the usual operations is a field.

(2.) The set $\mathbb{R}$ of real numbers with the usual operations is a field.

(3.) The set $\mathbb{Z}$ of integers with the usual operations is *not* a field. Why not?

(4.) The set $\mathbb{C}$ of complex numbers with the usual operations is a field.

(5.) Let $F=\{a,b\}$. Define the operations $+,\cdot$ on F as follows:

$+$	a	b
a	a	b
b	b	a

$\cdot$	a	b
a	a	a
b	a	b

It is easy to verify that $+,\cdot$ are commutative and associative.

For example, $a+b=b=b+a$, $a+(a+b)=a+b=b=a+b=(a+a)+b$.

We see that a behaves like an additive identity, in that $a+x=x$ for all $x\in F$. Furthermore, b behaves like a multiplicative identity, in that $b\cdot x=x$ for all $x\in F$. Let's therefore make a the designated element 0, and make b the designated element 1, so that $F=\{0,1\}$. Now our tables look like:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

To make F into a field, we must also see if we can define two unary operations $-$ and $^{-1}$. Since $1+1=0$ and $0+0=0$, we define $-: F \to F$ by: $-1=1$, $-0=0$. The operation $^{-1}$ needs only be defined for the element 1. As $1 \cdot 1 = 1$, we define $1^{-1}=1$. Thus

x	$-x$
0	0
1	1

x	x^{-1}
1	1

It is easy to see that the addition and multiplication defined on F are just ordinary division and multiplication, modulo 2.
The field F goes by the name $\mathbb{Z}_2$.

Next, we prove some basic results about arithmetic inside a field. Most of these *look* obvious, but that's only because they are already so familiar. To *formally* prove these results, we are allowed to use only the field axioms. Consequently, these results will be true in *any* field — not just the familiar ones, such as $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, but also the as yet unfamiliar ones, such as $\mathbb{Z}_p$ (p a prime).

Proposition 1.2.5 *The axioms* C^+, A^+, Id^+, Inv^+ *imply the following statements:*

(a) If $x+y=x+z$, then $y=z$; (cancellation)

(b) If $x+y=x$, then $y=0$; (uniqueness of additive identity)

(c) If $x+y=0$, then $y=-x$; (uniqueness of additive inverse)

(d) $-(-x)=x$;

Proof: (a) Suppose that $x+y=x+z$. Then

$$\begin{aligned} y &= 0+y && \mathrm{Id}^+ \\ &= (-x+x)+y && \mathrm{Inv}^+ \\ &= -x+(x+y) && \mathrm{A}^+ \\ &= -x+(x+z) && \text{assumption} \\ &= (-x+x)+z && \mathrm{A}^+ \\ &= 0+z && \mathrm{Inv}^+ \\ &= z && \mathrm{Id}^+ \end{aligned}$$

(b) If $x+y=x$, then $x+y=x+0$, so the result follows from (a).
(c) If $x+y=0$, then

$$\begin{aligned} y &= 0+y \\ &= (-x+x)+y \\ &= -x+(x+y) \\ &= -x+0 \\ &= -x \end{aligned}$$

(d) $-x + x = 0$, so by (c), we must have $x = -(-x)$.

⊣

Proposition 1.2.6 *The axioms* C$^{\cdot}$, A$^{\cdot}$, Id$^{\cdot}$, Inv$^{\cdot}$ *imply the following statements:*

(a) If $xy = xz$, *and* $x \neq 0$, *then* $y = z$; (cancellation)

(b) If $xy = x$, *and* $x \neq 0$, *then* $y = 1$; (uniqueness of multiplicative identity)

(c) If $xy = 1$, *and* $x \neq 0$, *then* $y = x^{-1}$; (uniqueness of multiplicative inverse)

(d) If $x \neq 0$, *then* $(x^{-1})^{-1} = x$;

Exercise 1.2.7 Prove the preceding proposition.

□

Note that the distributive law (D) was not used in proving the above statements. If we invoke the distributive law, we can prove more:

Proposition 1.2.8 *The field axioms imply the following statements:*

(a) $0 \cdot x = 0$;

(b) If $x \neq 0$, $y \neq 0$, *then* $xy \neq 0$; *thus if* $xy = 0$, *then either* $x = 0$ *or* $y = 0$;

(c) $(-x)y = -xy = x(-y)$;

(d) $(-x)(-y) = xy$;

Exercise 1.2.9 Prove the preceding proposition. Here are some hints:
(a) Justify the following string of equalities: $x + 0 \cdot x = 1 \cdot x + 0 \cdot x = (1 + 0) \cdot x = 1 \cdot x = x$. Now use Proposition 1.2.5(b).
(b) If x, y are non–zero, then $(x^{-1}y^{-1})(xy) = 1$. Hence, by (a), $xy \neq 0$. (Why can't we have $0 = 1$?)
(c) $(-x)y + xy = (-x + x)y = 0$, so $(-x)y = -(xy)$, by Proposition 1.2.5(c).
(d) Apply (c) twice, and invoke Proposition 1.2.5(d).

□

In this section, we concentrated on the *arithmetic* of fields. We saw that there are many different of fields, some finite (e.g. $\mathbb{Z}_2$) and some infinite (e.g. $\mathbb{C}$). The field axioms are therefore *insufficiently strong* to characterize the system of real numbers. In the next section, we go some way towards remedying this situation, by adding more axioms.

1.3 Ordered Fields

In addition to basic arithmetic, intuition about real numbers also contains the notion of order, i.e. we consider some real numbers to be less than others. This notion does not make sense in all fields. For example, the field $\mathbb{C}$ of complex numbers has no natural ordering. (Is $2 + 3i$

greater or less than $3 + 2i$? — the question does not make sense.) So the notion of order is something *extra*, something outside the arithmetic of fields. We must therefore write down a set of axioms for the behaviour of the order relation on the reals.

Recall that a *partial ordering* $\leq$ on a set is a binary relation satisfying the following axioms:

(PO-R) $s \leq s$
(PO-A) $s \leq t$ and $t \leq s$ imply $s = t$
(PO-T) $s \leq t$ and $t \leq u$ implies $s \leq u$

An additional axiom, (TO), strengthens the notion of partial ordering to that of a *total ordering*:

(TO) Either $s \leq t$ or $t \leq s$ (or both)

A remark on notation: We write "$s < t$" instead of "$s \leq t$ and $s \neq t$". Similarly "$s \geq t$" is just another way of saying "$t \leq s$". An analogous statement holds for "$s > t$".

An ordered field is a field which is also a totally ordered set, subject to two additional conditions:

Definition 1.3.1 An *ordered field* is a tuple $\langle F, +, \cdot, -, ^{-1}, 0, 1, \leq \rangle$ satisfying the field axioms (C^+), (A^+), (Id^+), (Inv^+), ($C^\cdot$), ($A^\cdot$), ($Id^\cdot$), ($Inv^\cdot$), (D), and the order axioms (PO-R), (PO-A), (PO-T), (TO), such that, in addition

(OF^+) $\forall x \forall y \forall z [x \leq y \to x + z \leq y + z]$
($OF^\cdot$) $\forall x \forall y [x > 0 \wedge y > 0 \to xy > 0]$

The fields $\mathbb{Q}$ and $\mathbb{R}$ are ordered fields (with the usual ordering). However,

- No finite field is an ordered field.
- The field $\mathbb{C}$ cannot be made into an ordered field.

You will be required to prove these facts soon.

Definition 1.3.2 If $x > 0$, we say that x is *positive*; if $x < 0$, we say that x is *negative*. We say that x, y have *opposite signs* if one of x, y is positive and the other negative.

The following proposition contains some familiar results on the interaction between order and arithmetic. Once again, we stress that these will hold true in *any* ordered field (not just the reals), as only the ordered field axioms will be used in the proof.

Proposition 1.3.3 *Let $\langle F, +, \cdot, -, ^{-1}, 0, 1, \leq \rangle$ be an ordered field.*

(a) $x < y$ if and only if $y - x > 0$;

(b) If $x \neq 0$, then x and $-x$ have opposite signs.

(c) If $x > 0$ and $y < z$, then $xy < xz$; if $x < 0$ and $y < z$, then $xy > xz$;

(d) If $x \neq 0$, then $x^2 > 0$;

(e) $1 > 0$;

(f) If $x \neq 0, y \neq 0$, then x, y have opposite signs if and only if $xy < 0$.

(g) $x > 0$ implies $x^{-1} > 0$; $x < 0$ implies $x^{-1} < 0$;

(h) If $0 < x < y$, then $0 < y^{-1} < x^{-1}$;

Proof: (a) $x < y$ implies $x + (-x) < y + (-x)$ by (OF^+), so $0 < y - x$. Similarly $0 < y - x$ implies $x + 0 < x + (y - x)$.
(b) If $x < 0$, then $x + (-x) < 0 + (-x)$, so $0 < -x$. A similar proof works for the case $x > 0$.
(c) $x > 0$ and $z - y > 0$ implies $x(z - y) > 0$ by $(\mathrm{OF}^\cdot)$. A similar proof works for the other case.
(d) This is clear if $x > 0$. Else, $-x > 0$ (why?), and so $(-x) \cdot (-x) > 0$, by $(\mathrm{OF}^\cdot)$. But $(-x) \cdot (-x) = x^2$, by Proposition 1.2.8(d).
(e) By (d);
(f) If $x > 0, y < 0$, then $xy < 0$, by (c). Conversely, suppose that $xy < 0$. Then $x \neq 0$ and $y \neq 0$. Now if x, y are both positive, then xy is positive, by $(\mathrm{OF}^\cdot)$; if x, y are both negative, then $-x, -y$ are both positive (by (b)), so that $(-x)(-y) = xy$ is positive, by $(\mathrm{OF}^\cdot)$.
(g) Since $xx^{-1} = 1 > 0$, x and x^{-1} cannot have opposite signs.
(h) Suppose that $0 < x < y$. Then $x^{-1}, y^{-1} > 0$, so $0 < x \cdot x^{-1} < yx^{-1}$, by $(\mathrm{OF}^\cdot)$. Hence $1 < yx^{-1}$, and so $y^{-1} < y^{-1}(yx^{-1})$.

⊣

Exercise 1.3.4 (a) Suppose that $\langle F, +, \cdot, -, ^{-1}, 0, 1\rangle$ is a field, and that $P \subseteq F$ has the following properties:

(i) For each $x \in F$, exactly *one* of the following is true:

$$x = 0 \quad \text{or} \quad x \in P \quad \text{or} \quad -x \in P$$

(ii) $x, y \in P$ implies $xy \in P$

(iii) $x, y \in P$ implies $x + y \in P$;

Define a binary relation $\leq$ on F by

$$x \leq y \Leftrightarrow x = y \text{ or } y - x \in P$$

Show that $\leq$ makes F into an ordered field. Also show that P is precisely the set of positive elements.

(b) Prove that there are no finite ordered fields.
[Hint: Show that

$$0 < 1 < 1 + 1 < 1 + 1 + 1 < 1 + 1 + 1 + 1 < \ldots$$

]

(c) Show that $\mathbb{C}$ cannot be made into an ordered field — it is impossible to define a total ordering $\leq$ on $\mathbb{C}$ so that (OF^+) and $(\mathrm{OF}^\cdot)$ are satisfied.

□

1.4 The Continuum

Let's take stock for a moment: We are trying to find a complete set of axioms for the real numbers, i.e we are attempting find a set of rules which completely capture our intuition about the behaviour of the reals. Our intuition involves both arithmetic and order–theoretic properties, and, so far, we've written down 15 axioms, the axioms of an ordered field: (C^+), (A^+), (Id^+), (Inv^+), ($C^\cdot$), ($A^\cdot$), ($Id^\cdot$), ($Inv^\cdot$), (D),(PO-R), (PO-A), (PO-T), (TO), (OF^+), and ($OF^\cdot$). Using these axioms, *and nothing else*, we've managed to prove a number of interesting properties: $(-x)(-y) = xy$; squares (x^2) are non–negative, etc. These properties hold in any ordered field.

So do the ordered field axioms completely capture our intuition about the behaviour of the reals? No. The set $\mathbb{Q}$ of rational numbers also forms an ordered field, and we know that $\mathbb{Q} \neq \mathbb{R}$ (e.g. there are numbers, such as $\sqrt{2}$, which are not rational). Thus we have some additional intuition which allows us to distinguish the set of reals from the set of rationals. What could it be?

Geometry comes into play. We have an additional intuition about non–negative real numbers as being *lengths* of straight line segments: We can measure the length of a line segment using a ruler, and the length will be a real number. In this way, we come to regard the set of real numbers as points on a straight line which extends indefinitely in both directions.

Example 1.4.1 (1) Consider an isosceles right–angled triangle, with right–angle sides both 1 unit in length. Our intuition dictates that the hypotenuse have a length. By Pythagoras' Theorem, the length of the hypotenuse is a number x satisfying $x^2 = 2, x \geq 0$.

(2) Consider the graph of the parabola $f(x) = x^2 - 2$ in the Cartesian plane. We see that $f(0) < 0$, and that $f(2) > 0$. Our intuition dictates that the graph cut the x–axis *somewhere* between 0 and 2. It is cut at an x satisfying $x^2 = 2, x \geq 0$.

□

Of course, you say, in both examples we are seeking the number $x = \sqrt{2}$. However, the symbol $\sqrt{}$ is not (yet) part of our language. The existence of roots is something genuinely new, as we shall show in the next section.

The following proposition should be familiar to you:

Proposition 1.4.2 *There is no rational number $x \in \mathbb{Q}$ such that $x^2 = 2$.*

Proof: The proof is by contradiction. Suppose that $x^2 = 2$ and that $x = \frac{n}{m}$. Choose m_0 to be the *least* positive integer such that there is an integer n_0 for which $x = \frac{n_0}{m_0}$. Clearly

$$2m_0^2 = n_0^2$$

so that n_0 is even, i.e. $n_0 = 2l_0$ for some integer l_0. Then

$$m_0^2 = 2l_0^2$$

so m_0 is even, i.e. $m_0 = 2k_0$ for some positive integer k_0. It follows that $x = \frac{l_0}{k_0}$, in contradiction to the choice of m_0 (since $k_0 < m_0$).

⊣

So $\sqrt{2}$, if it exists, is irrational.

Exercise 1.4.3 (a) Let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an n^{th}–degree polynomial with integer coefficients $a_0, \ldots, a_n$, where $a_n \neq 0$. Suppose that $\frac{p}{q} \in \mathbb{Q}$ is a root of $f(x)$, where $p, q \in \mathbb{Z}$ are relatively prime[1]. Show that p is a factor of a_0 and that q is a factor of a_n.

(b) Consider the special case where $a_n = 1$. Show that all *rational* roots of $f(x)$ must be integers.

(c) Conclude that the following numbers, if they exist, are irrational: $\sqrt{2}, \sqrt[3]{12}, (5 - \sqrt{2})^{\frac{1}{3}}$.

□

Here's a brief summary of the salient points of this section:

- We have provided a set of axioms that capture our intuition about the arithmetic and order–theoretic properties of real numbers.
- We also have a geometric intuition about non–negative real numbers as being the lengths of line segments. Applying the Theorem of Pythagoras to the hypotenuse of an isosceles right–angled triangle, we are lead to believe that there exists a non–negative real number x with the property that $x^2 = 2$.
- However, we proved that x, if it exists, cannot be a rational number.
- Because the field $\mathbb{Q}$ of rational numbers satisfies all the axioms proposed, it follows that the existence of x cannot be proved from those axioms alone.
- Thus the set of axioms proposed so far is inadequate: It does not completely capture our intuition about the real number system, because there are "truths" that we cannot prove.
- It is therefore necessary to find at least one more axiom.

From where will we get such an axiom? The intuition that leads us to believe in the existence of the number $\sqrt{2}$ has its origins in the geometric concept of length. Thus far, we have not considered any geometric axioms at all. We would like to keep things as simple as possible, and avoid an axiomatization that depends too much on geometric concepts. Is it really necessary to incorporate, say, Euclid's axioms of plane geometry, to formalize the notion of the length of a line segment and the notion of a right–angled triangle, all in order to prove Pythagoras' Theorem, so that we can finally obtain the existence of $\sqrt{2}$?

Fortunately not. Perform the following thought experiment, an exercise in visualization: Let N be a natural number, and consider an arbitrary, but non–empty, set $\mathcal{L} = \{L_i : i \in I\}$ of line segments, each of length $\leq N$. Align them, and stack them together on top of each other, so that you see just a *single* line segment L. Clearly the following hold:

(i) The length of the stack, L, is greater than or equal to the length of each $L_i \in \mathcal{L}$;

[1] Two integers are relatively prime if their greatest common divisor is 1. In this case, this means that the fraction $\frac{p}{q}$ cannot be simplified any further.

(ii) The length of L is finite, being less than or equal to N.

(iii) Any line segment which is strictly shorter than L is also shorter than some $L_i \in \mathcal{L}$.

(iv) Thus L is the *shortest* line segment which has length greater than or equal to the length of each $L_i \in \mathcal{L}$.

Note that L may be strictly longer than each of the L_i. For example, if the index set is the set of natural numbers (i.e. $I = \mathbb{N}$) and $L_i = 1 - 2^{-i}$, then each $L_i < 1$. But the length of the stack is $L = 1$. Thus the length L need not be a member of the given set of lengths $\{L_i : i \in I\}$. It is something *new*.

We now rephrase and generalize the above intuition as follows:

> *If $A = \{a_i : i \in I\}$ is a non–empty set of real numbers which is bounded above, then it has a* least *upper bound, i.e. there exists a number a such that $a \geq a_i$ for all $a_i \in A$, and if $a' < a$, then there is $a_i \in A$ such that also $a' < a_i$.*

It turns out that with this final axiom we have completely characterized the set of real numbers.

1.5 The Completeness Axiom

We begin with some definitions:

Definition 1.5.1 Let $(P, \leq)$ be a total ordering[2] and let $A \subseteq P$.

(a) We say that an element $u \in P$ is an *upper bound* for A if and only if

$$\forall a \in A(a \leq u)$$

(b) Similarly, we say that $l \in P$ is an *lower bound* for A if and only if

$$\forall a \in A(l \leq a)$$

(c) We say that A is *bounded* if and only if it has both an upper bound and a lower bound.

(d) We say that u_0 is the *supremum*, or *least upper bound* of A if and only if the following hold:

(i) u_0 is an upper bound of A;

(ii) If u is any upper bound of A, then $u_0 \leq u$.

We write

$$u_0 = \sup A \quad \text{or} \quad u_0 = \text{l.u.b.}(A)$$

(e) We say that l_0 is the *infimum*, or *greatest lower bound* of A if and only if the following hold:

[2] i.e. $(P, \leq)$ satisfies (PO-R), (PO-A), (PO-T) and (TO).

(i) l_0 is a lower bound of A;

(ii) If l is any lower bound of A, then $l \leq l_0$.

We write

$$l_0 = \inf A \quad \text{or} \quad l_0 = \text{g.l.b.}(A)$$

(f) We say that u_0 is the *maximum* of A, and write $u_0 = \max A$, if and only if

$$u_0 \in A \quad \text{and} \quad u_0 = \sup A$$

(g) Similarly, we say that l_0 is the *minimum* of A, denoted $l_0 = \min A$, if and only if

$$l_0 \in A \quad \text{and} \quad l_0 = \inf A$$

□

Definition 1.5.2 Let $(P, \leq)$ be a total ordering.

(a) Let $a, b \in P$ with $a \leq b$. We define the following sets:

$$[a, b] = \{x \in P : a \leq x \leq b\}$$
$$(a, b) = \{x \in P : a < x < b\}$$
$$(a, b] = \{x \in P : a < x \leq b\}$$
$$[a, b) = \{x \in P : a \leq x < b\}$$

(b) A set $A \subseteq P$ is called an *interval* (in P) if and only if whenever $a, b \in A$ with $a \leq b$, then $[a, b] \subseteq A$.

Examples 1.5.3 (1) If $a < b$ in a total ordering $(P, \leq)$, then $[a, b], (a, b), (a, b], [a, b)$ are intervals.

(2) A subset A of a total ordering $(P, \leq)$ is bounded if and only if there exist $a, b \in P$ such that $A \subseteq [a, b]$.

(3) In $(\mathbb{R}, \leq)$, we have:

(a) $1 = \max[0, 1] = \sup[0, 1]; \quad 0 = \min[0, 1] = \inf[0, 1]$

(b) $1 = \sup[0, 1)$, but $\max[0, 1)$ does not exist.

(c) If $A = \{x \in \mathbb{R} : x^2 \leq 2\}$, then $\sup A = \sqrt{2}$.

(d) If $A \neq \emptyset$, then $\inf A \leq \sup A$.

(e) If $A = \emptyset$, then A has no supremum and no infimum. First note that every $u \in \mathbb{R}$ is an upper bound for A — for if u is not an upper bound, then there must be an $a \in A$ such that $u < a$. However, A is empty, so no such a can be found.

Similarly, every real number is also a lower bound for A. We therefore write

$$\sup \emptyset = -\infty \qquad \inf \emptyset = +\infty$$

(4) In $(\mathbb{Q}, \leq)$, the set $A = \{x \in \mathbb{Q} : x^2 \leq 2\}$ has no supremum: If $u \in \mathbb{Q}$ is an upper bound for A, then u is a rational number which is greater than $\sqrt{2}$. We can then choose a rational number u' such that $\sqrt{2} < u' < u$. Thus: given any upper bound u for A, we can find another upper bound $u' < u$. Hence A has no least upper bound (in $\mathbb{Q}$).

□

Definition 1.5.4 (Completeness)
Let F be an ordered field. We say that F is *complete* if and only whenever a non–empty $A \subseteq F$ has an upper bound, then it has a least upper bound, i.e. $\sup A$ exists (and belongs to F).

Exercise 1.5.5 Show that if an ordered field F is complete, then any non–empty subset of F which is bounded below has a greatest lower bound. Thus: In a complete ordered field any non-empty bounded set has both a supremum and an infimum.

□

Note that $\mathbb{Q}$ is not complete: The set $A = \{x \in \mathbb{Q} : x^2 \leq 2\}$ has an upper bound in $\mathbb{Q}$, but no supremum. However, our intuition about real numbers as lengths of line segments led us to conclude, via an experiment in visualization, that the set of real numbers is complete. This is our final axiom for the system of real numbers. The Completeness Axiom allows us to distinguish between $\mathbb{R}$ and $\mathbb{Q}$.

We state the following result without proof:

Theorem 1.5.6 *There exists a complete ordered field* $(\mathbb{R}, +, \cdot, -, ^{-1}, 0, 1, \leq)$. *Moreover, there is essential only one complete ordered field, in the sense that any two complete ordered fields are* isomorphic.

Remarks 1.5.7 The notion of *field isomorphism* will be studied in detail in an algebra course, so we won't give a formal definition here. To say that two algebraic structures, such as two fields F and G, are *isomorphic*, means that they are essentially the same object.

As an analogy, consider two copies of *War and Peace*. They are not *equal*: One is in my left hand, the other in my right hand; one is a hardback, the other a paperback, etc. But they are nevertheless the *same* book. Similarly, two mathematical structures can be the *same*, without being *equal*.

More formally, two fields F, G are isomorphic if there is a bijection $\varphi : F \to G$ which *preserves* all the operations. Thus F and G have the same structure, and it is only the *names* f, g of the field elements that are different. The element f of F corresponds to the element $\varphi(f)$ of G. Any relationship between elements $f_1, \ldots, f_n$ of F will also hold for $\varphi(f_1), \ldots, \varphi(f_n)$ in G. In particular, if the relation $f_1 + f_2 = f_3$ holds in F, then the relation $\varphi(f_1) + \varphi(f_2) = \varphi(f_3)$ holds in G, so that $\varphi(f_1 + f_2) = \varphi(f_1) + \varphi(f_3)$. Similarly, $\varphi(f^{-1}) = \varphi(f)^{-1}$, $\varphi(0) = 0$, etc. (Note that the $+$ in $\varphi(f_1 + f_2)$ refers to addition in F, whereas the $+$ in $\varphi(f_1) + \varphi(f_2)$ refers to addition in G. The same goes for the other operations and constants.)

□

Exercise 1.5.8 Consider the following two–player game, called FIFTEEN: Players take turns to pick a number from the set $\{1, 2, \ldots, 9\}$. No number may be picked more than once (i.e. if Player I picks the number 3, that cannot be picked later by either Player I or Player II) — perhaps the simplest way would be to play would be with 10 cards, an ace (with value 1), 2, 3, ..., 9, with players alternating to picking a card . The object is to be the first player to possess three numbers (cards) whose values add up to 15.

Convince yourself that this game is, in a sense, isomorphic to the game of *noughts–and–crosses* (also known as *tic–tac–toe*).

6	7	2
1	5	9
8	3	4

Now astound friends and family with your mental skills by persuading them to play FIFTEEN, while you secretly play noughts–and–crosses.

(The point is that, even when "things" are "the same", one *representation* of that "thing" may be vastly superior to another.)

□

We have now accomplished our goal: We *define* the system of real numbers to be a complete ordered field. By the above theorem, there is essentially only one such structure. Thus, whenever we say "such and such is true in the set of real numbers", we mean "such and such is true in a complete ordered field."

Remarks 1.5.9 (a) So now we have *defined* the reals to be a complete ordered field. But, I hear you mutter, isn't a real number an object like 3.14159265...? Not quite.

3.14159265... (decimal) is a *representation* of a certain real number. The symbol π is another, as is 11.001001000011111... (in binary). All of these represent the same real number, but they are not the same as that number. We will prove in the next chapter that every element of a complete ordered field has a decimal (or binary) representation.

Note that we neatly side–step the following question: "What is a real number?" This question is not answered by the proof of Theorem 1.5.6, as there are several proofs, each giving a different construction of real numbers.

The question is not important for the purpose of mathematics: It is the *relationships* between real numbers that count, not their "true nature" (whatever *that* might mean).

(b) Note that the Completeness Axiom is of a very different nature to the other axioms. All the other axioms speak about elements of fields. For example, for every $x, y, z \in F$, the elements ($x + y_+ z$ and $x + (y + z)$ are equal (associativity of $+$); if $x \leq y$ and $y \leq z$ then $x \leq z$ (transitivity of '), etc. In logical parlance, these are *first order* axioms.

The Completeness axiom, on the other hand speaks about sets of field elements: For every $A \subseteq F$, if A is bounded above, then X has a supremum. In logical parlance, this is a *second order* axiom.

□

So far, while operating at the intuitive level, we have always assumed that $\mathbb{Q} \subseteq \mathbb{R}$. But now we have formally defined $\mathbb{R}$ as a complete ordered field, we must ensure that this is true. The proof requires the notion of field isomorphism, and therefore properly belongs to algebra. We merely provide an intuitive outline:

Proposition 1.5.10 *The field $\mathbb{Q}$ of rational numbers is the* smallest *ordered field, in the following sense: Any ordered field contains (a copy of) $\mathbb{Q}$ as a subfield. In particular, $\mathbb{Q} \subseteq \mathbb{R}$.*

"Proof": Let F be an ordered field. We show that each rational number $q \in \mathbb{Q}$ can be identified with an element $(q)_F \in F$ If $n \in \mathbb{N}$, we identify it with the field element

$$(n)_F = \underbrace{(1 + 1 + \cdots + 1)}_{n \text{ times}} \in F$$

Also identify the number zero with the zero field element, i.e. $(0)_F = 0$. If n is a negative integer, we define $(n)_F = -(-n)_F$. ($-n$ is a positive integer, so $(-n)_F$ has already been defined.). Next, if $\frac{n}{m} \in \mathbb{Q}$, we identify it with the F–element

$$\left(\frac{n}{m}\right)_F = (n)_F (m)_F^{-1} \qquad \text{for } m \neq 0$$

It is not hard to show that the map $\varphi : \mathbb{Q} \to \mathbb{F} : \frac{n}{n} \mapsto (\frac{n}{m})_F$ is well-defined and injective, and that it preserves the relations between elements, i.e. that φ is an isomorphism between $\mathbb{Q}$ and the subfield $\{(\frac{n}{m})_F : m, n \in \mathbb{Z}, m \neq 0\}$.

⊣

Theorem 1.5.11 *Let F be a complete ordered field.*

(a) F satisfies the archimedean property*: For any $x > 0$ and any y in F there exists $n \in \mathbb{N}$ such that $nx > y$.*

(b) The field $\mathbb{Q}$ of rationals is dense in F: whenever $x < y$ in F there is $q \in \mathbb{Q}$ such that $x < q < y$.

Proof: (a) Let $A = \{nx : n \in \mathbb{N}\}$. Then $A \neq \emptyset$. If there is no $n \in \mathbb{N}$ such that $nx > y$, then y is an upper bound for A. Thus the completeness axiom guarantees that $a_0 = \sup A$ exists (in F). Now $a_0 - x < a_0$, as $x > 0$, so $a_0 - x$ is not an upper bound of A. Hence there exists $m \in \mathbb{N}$ such that $a_0 - x < mx$. Then $a_0 < (m+1)x$. But $(m+1)x \in A$, and a_0 is an upper bound for A — contradiction.
(b) Recall that any ordered field contains (a copy of) the field $\mathbb{Q}$, by Proposition 1.5.10. We shall show that there exist $m \in \mathbb{Z}, n \in \mathbb{N}$ such that (regarded as a member of F), $x < \frac{m}{n} < y$.

Now if $x < y$, then $y - x > 0$, and so the archimedean property allows us to find a non–negative integer n such that $n(y-x) > 1$. Similarly, there are non–negative integers m_1, m_2 such that $m_1 > nx, m_2 > -nx$ — just consider the two cases $x \geq 0, x < 0$. It follows that $-m_2 < nx < m_1$, and so there is a smallest integer m such that $nx < m$. It follows that $m - 1 \leq nx$, and so

$$nx < m \leq 1 + nx < ny$$

which yields $x < \frac{m}{n} < y$. Division by n is possible, because $n > 0$.

⊣

How did we get to the completeness axiom? Our intuition about real numbers as lengths led us to believe in the existence of a number x such that $x^2 = 2$. We subsequently found that such an x could not be rational, and we concluded that we could not deduce the existence of x from the ordered field axioms alone. We then performed a experiment in visualization, stripped it of its geometric content, and wrote down the completeness axiom.

So can we prove the existence of $\sqrt{2}$ from just the ordered field axioms and the completeness axiom? Indeed, we can:

Proposition 1.5.12 *Let F be a complete ordered field, let $n \in \mathbb{N}$, and let $x > 0$ in F. Then there exists a unique $y \in F$ such that $y > 0$ and $y^n = x$. We denote this y by $y = \sqrt[n]{x}$.*

Exercise 1.5.13 We prove Proposition 1.5.12.

(a) First show that there is a unique such y, i.e. that if $y_1, y_2 > 0$ are such that $y_1^n = x = y_2^n$, then $y_1 = y_2$.

(b) Define $A = \{t \in F : t > 0, t^n \leq x\}$. Show that $x(x+1)^{-1} \in A$.

(c) Show that $1 + x$ is an upper bound of A.

(d) Thus A is non–empty and bounded above. Let $y = \sup A$. We shall show that $y^n = x$. To do so we shall need the following inequality: For $0 < a < b$ in F, we have

$$b^n - a^n < (b-a)nb^{n-1}$$

Use the identity $b^n - a^n = (b-a)(b^{n-1} + b^{n-2}a + \cdots + ba^{n-2} + a^{n-1})$ to prove that this inequality holds.

(e) We apply this inequality twice — once to show that we cannot have $y^n > x$, and once to show that we cannot have $y^n < x$.

Suppose $y^n > x$. Define

$$k = \frac{y^n - x}{ny^{n-1}}$$

Note that $0 < k < \frac{y}{n} \leq y$. Now show that if $t = y - k$ then $y^n - t^n \leq y^n - x$, and deduce that $t^n > x$.

(f) Thus t is an upper bound of A. Hence explain why assuming that $y^n > x$ leads to contradiction.

(g) Next suppose that $y^n < x$. Explain why we may choose an $h \in F$ such that $0 < h < \min(\frac{x-y^n}{n(y+1)^{n-1}}, 1)$.

(h) Now define $s = y + h$, and show that

$$s^n - y^n < hns^{n-1} < hn(y+1)^{n-1} < x - y^n$$

Conclude that $s^n < x$.

(i) Explain why assuming that $y^n < x$ leads to contradiction.

(j) Conclude that $y^n = x$, as required.

□

Exercise 1.5.14 Let F be an ordered field $a < b$ in F. We have seen that sets of the form $[a, b], (a, b), (a, b], [a, b)$ are bounded intervals.

(a) Show that not every bounded non–empty interval need be of this form.
[Hint: Consider the set $A = \mathbb{Q} \cap (0, \sqrt{2})$. This is a bounded interval in the field $\mathbb{Q}$. However, there are no a, b in $\mathbb{Q}$ (!!) such that $A = (a, b)$ or $A = (a, b]$, etc.]

(b) Show that if F is a *complete* ordered field, then every bounded non–empty interval must be of this form.

□

Exercise 1.5.15 Let F be a complete ordered field, and let $n \in \mathbb{N}$. If $x, y > 0$ in F, show that $(xy)^{\frac{1}{n}} = x^{\frac{1}{n}} y^{\frac{1}{n}}$.

□

The following exercise proves an important property.

Exercise 1.5.16 (Nested Interval Property)
A family $\mathcal{A} = \{A_i : i \in I\}$ be a family of subintervals of an ordered field is said to be *nested* if and only if it satisfies the following condition: Whenever $i, j \in I$, either $A_i \subseteq A_j$ or $A_j \subseteq A_i$.

(a) Prove that $\mathbb{R}$ has the nested interval property: Any nested family of *closed* intervals has non–empty intersection.

(b) Show that the field $\mathbb{Q}$ does not have the nested interval property.

[Hint: (a) Let A_i be a closed interval with endpoints a_i and b_i. Show that for all $i, j \in I$ we have $a_i \leq b_j$. Conclude that $\{a_i : i \in I\}$ has an upper bound.
(b) Choose rational $a_i < b_i$ such that a_i, b_i converge to the same irrational number.]

□

Chapter 2

Sequences and Series

In this chapter we start doing proper analysis. We begin our attack on the notion of *limit* by discussing limits of sequences.

We define $\mathbb{N}$ to be the set $\{1, 2, 3, \dots\}$ (excluding 0).

2.1 Introduction

Intuitively, a sequence in a set X is a list

$$x_1, x_2, x_3, \dots, x_n, \dots$$

of members of X. Now we can think of such a sequence as a *function* f which assigns to each $n \in \mathbb{N}$ an element x_n of X. Thus $f(1) = x_1, f(2) = x_2$, etc. We turn this insight into a formal definition:

Definition 2.1.1 A sequence in $\mathbb{R}$ is a *function* $f : \mathbb{N} \to \mathbb{R}$.
We will write $\langle x_n \rangle_n$, or $\langle x_n : n \in \mathbb{N} \rangle$, or $\langle x_n \rangle_{n=1}^{\infty}$ for the function f with the property that $f(n) = x_n$ (for all $n \in \mathbb{N}$).

Remarks 2.1.2 Often we may want to consider a sequence of the form $x_0, x_1, x_2, \dots$, and sometimes even of the form $x_{-2}, x_{-1}, x_0, x_1, \dots$. Each of these can be thought of as a function. The first sequence is a function with domain $\{0, 1, 2, \dots\}$, and the second has domain $\{-2, -1, 0, 1, \dots\}$. We will write these as $\langle x_n \rangle_{n=0}^{\infty}$ and $\langle x_n \rangle_{n=-2}^{\infty}$ respectively.

□

Examples 2.1.3 Here are some examples of sequences in $\mathbb{R}$:

(a) $\langle (-1)^n \rangle_n$ is the sequence $-1, 1, -1, 1 \dots$, i.e. alternately -1 and 1.

(b) $\langle \frac{1}{n} \rangle_n = 1, \frac{1}{2}, \frac{1}{3}, \dots$.

(c) We may also define a sequence *inductively* (or *recursively*), e.g. if $x_1 = x_2 = 1$ and $x_{n+2} = x_n + x_{n+1}$ for $n \in \mathbb{N}$, then

$$\langle x_n \rangle_n = 1, 1, 2, 3, 5, 8, \dots$$

which is better known as the *Fibonacci* sequence.

□

Exercises 2.1.4 1. Write out the first four terms of the following sequences:

(a) $\langle 2^n \rangle_{n=0}^{\infty}$

(b) $\langle x_n \rangle_{n=-1}^{\infty}$, where $x_n = \frac{3n+1}{|4n-1|}$.

2. Given a real number x, let $g(x)$ be the greatest integer which is $\leq x$. Given that $x = 3.14159265\ldots$, write out the first nine terms of the sequence defined inductively by

$$x_0 = g(x) \qquad x_n = g\big(10^n(x - \sum_{k=0}^{n-1} 10^{-k} x_k)\big) \quad \text{for } n \geq 1$$

3. Find a general formula for x_n as a function of n, given that

(a) $x_1 = 1$, $x_{n+1} = x_n + 2$.

(b) $x_1 = 1, x_{n+1} = x_n + (n+1)$

□

You already have an intuitive understanding of the notion of convergence. For example,

$$1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4} \ldots \text{ converges to } 0$$
$$\tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \tfrac{4}{5} \ldots \text{ converges to } 1$$
$$3, 3.1, 3.14, 3.141, 3.1415 \ldots \text{ converges to } \pi$$

Intuitively, if we say that a sequence $\langle x_n \rangle_n$ converges to x, we mean that the elements of the sequence lie closer and closer to x. Still, it's not quite that straightforward; we must be careful:

(i) The terms of $1, 1.2, 1.22, 1.222, \ldots$ lie closer and closer to 3: The 2^{nd} term lies closer to 3 than the 1^{st}, and the 3^{rd} term lies closer to 3, than the 2^{nd}, etc.
Nevertheless, the sequence does not converge to 3 (but to $\frac{11}{9}$).
You might say that the problem is that the terms *never get to* 3. True, but they don't ever quite reach $\frac{11}{9}$ either...

(ii) The terms of the sequence $1, \frac{1}{2}, 1, \frac{3}{4}, 1, \frac{7}{8}, 1, \frac{15}{16}, \ldots$ do not lie "*closer and closer*" to 1: The 1^{st} term is closer to 1 than the 2^{nd}, the 99^{th} term is closer than the 100^{th} (and also than the 10 billionth), etc.
Thus some of the later terms lie further away from 1 than some of the earlier terms.
Yet the sequence certainly converges to 1.

It is therefore clear that, in defining the notion of convergence, we must be more precise than "$\langle x_n \rangle_n$ *converges to* x *if and only if the terms* x_n *lie closer and closer to* x" — this is too ambiguous to be useful.

The next few sections are therefore devoted to analyzing exactly what we mean when we say that a sequence converges.

Exercises 2.1.5 You need not supply *formal* proofs for the following exercises, but do show any relevant calculations.

1. Determine whether or not the following sequences converge, and if it converges, write down the limit.
   (a) $x_n = \frac{n(n+3)}{2n(2n+2)}$
   (b) $x_n = \frac{2^n}{n!}$
   (c) $x_n = 3 + (-1)^n$
   (d) $x_n = \sqrt{n+1} - \sqrt{n}$ [Hint: Show that $x_n = \frac{1}{\sqrt{n+1}+\sqrt{n}}$]
   (e) $x_n = \sqrt{n^2+n} - n$
2. Can a sequence of rational numbers converge to an irrational number? Can a sequence of irrational numbers converge to a rational number?

□

2.2 Definition of Convergence

2.2.1 "Infinitely Often" and "Eventually"

Here are two related ideas which will help elucidate the notion of convergence:

Let P be a property that a real number may (or may not) have. We write $P(x)$ if x has the property P. [For example, P could be the property of being positive, so that $P(1.23)$, but $\neg P(-\pi)$. Or Q could be the property of being irrational, in which case $\neg Q(1.23)$, whereas $Q(-\pi)$.] Now suppose that $\langle x_n \rangle_n$ is a sequence in $\mathbb{R}$, and that P is a property:

- We say $\langle x_n \rangle_n$ has property P *infinitely often* iff there are infinitely many n for such that $P(x_n)$ is true.
- We say $\langle x_n \rangle_n$ has property P*eventually* iff $P(x_n)$ is true for all n from some point onwards.

These are *intuitive* descriptions, not formal definitions — we'll come to that. But first, some examples:

Examples 2.2.1 (1) The sequence $1, 2, 3, 4, 5, \ldots$ is *prime* infinitely often: Infinitely many terms are prime numbers. It is also even infinitely often. It is eventually greater than 10^{10}.

(2) The sequence $-2, -1, 0, 1, 2, 3, \ldots$ is *positive* eventually: From the fourth term onwards, all terms are positive (i.e. > 0).

□

Exercises 2.2.2 (1) Let $x_n = \sin \frac{n\pi}{2}$. Show that $\langle x_n \rangle_n$ is strictly positive infinitely often, strictly negative infinitely often, and zero infinitely often.

(2) Let $x_n = 2 + \frac{(-1)^n}{n}$. Show that $|x_n - 2| < \frac{1}{1000}$ eventually.

(3) Let $x_n = 2 + \frac{(-1)^n}{n}$ for $n = 1, 2, 3, \ldots$. Let $P(x)$ be the property of "being $\geq x$". For which x does $P(x)$ hold (i) eventually, (ii) infinitely often?

□

Remarks 2.2.3

- First note that $\langle x_n \rangle_n$ has property P *eventually* if and only if there exists an $N \in \mathbb{N}$ such that every term after the N^{th} has property P. Formally,

$$(\exists N \in \mathbb{N})\ (\forall n \geq N)\ [x_n \text{ has property } P] \qquad (\dagger)$$

  We will take this to be a formal definition (cf. Definition 2.2.4)

- Thus $\langle x_n \rangle_n$ has property P eventually if and only if *all but finitely many* terms have property P (i.e. at most finitely many terms have property $\neg P$).

- Next note that if $\langle x_n \rangle_n$ has property P infinitely often, then the following is true:

  Given any natural number $N \in \mathbb{N}$, there is a natural number $n \geq N$ such that x_n satisfies property P, i.e.

$$(\forall N \in \mathbb{N})\ (\exists n \geq N)\ [x_n \text{ has property } P] \qquad (*)$$

  For if this is not the case, then there is some N such that *no* $n \geq N$ *has property* P. Thus if x_n *does* have property P, then $n < N$.
  But then there are only finitely many x_n which have property P — at most those x_n for $n = 1, 2, \ldots, N-1$!! This contradicts the assumption that $\langle x_n \rangle_n$ has property P *infinitely* often.

  It follows that if $\langle x_n \rangle_n$ has property P infinitely often, then $(*)$ is true.

- Conversely, if $(*)$ holds, i.e. if given any N there is a later $n \geq N$ such that x_n has property P, then there must be infinitely many elements which have property P:

  - Take $N := 0$: There must be an $n_1 \geq 0$ such that x_{n_1} has property P.
  - Then take $N := n_1 + 1$: There must be $n_2 > n_1$ such that x_{n_2} has property P.
  - Then take $N := n_2 + 1$ there must be an $n_3 > n_2 \ldots$, etc.

  We thus obtain an infinite sequence $n_1 < n_2 < n_3 < \ldots$ of natural numbers such that each x_{n_k} has property P.

  It follows that if $(*)$ is true, then $\langle x_n \rangle_n$ has property P infinitely often.

- Hence

  $(*)$ is equivalent to "$\langle x_n \rangle_n$ *has property* P *infinitely often*"

  We will therefore take $(*)$ to be a formal definition (cf. Definition 2.2.4).

- Further note that if $\langle x_n \rangle_n$ has property P eventually, then it also has property P infinitely often.

- Finally, observe that *infinitely often* and *eventually* are closely related: If it is not the case that a sequence $\langle x_n \rangle_n$ has property P infinitely often, then eventually $\langle x_n \rangle_n$ must have property $\neg P$.

$$\neg(\forall N)(\exists n \geq N)\ [x_n \text{ has property } P] \ \equiv\ (\exists N)(\forall n \geq N)\ [x_n \text{ has property } \neg P]$$

For example, if the sequence $\langle x_n \rangle$ is *not* positive infinitely often, then it has only finitely many positive terms. Thus from some point onwards, every term must be non–positive. Similarly, if a sequence does not have property P eventually iff it has property $\neg P$ infinitely often:

$$\neg(P \text{ eventually}) \equiv (\neg P \text{ infinitely often}) \qquad \neg(P \text{ infinitely often}) \equiv (\neg P \text{ eventually})$$

□

The above remarks contain a set of formal definitions:

Definition 2.2.4 (a) We say that a sequence $\langle x_n \rangle_n$ has property P *infinitely often* if and only if

$$(\forall N \in \mathbb{N})\ (\exists n \in \mathbb{N})\ [n \geq N\ \wedge\ x_n \text{ has property } P]$$

(b) We say that a sequence $\langle x_n \rangle_n$ has property P *eventually* if and only if

$$(\exists N \in \mathbb{N})\ (\forall n \in \mathbb{N})\ [n \geq N\ \rightarrow\ x_n \text{ has property } P]$$

2.2.2 Convergence to 0

We first define what it means for a sequence $\langle x_n \rangle_n$ of non–negative real numbers to converge to zero. Intuitively:

To say that $x_n \to 0$ means that $\langle x_n \rangle$ is "small" eventually, for any measure of "smallness".

The notion "small" is subjective, so we will demand that it holds for absolutely anybody's idea of "small". Specifically, suppose you define "small" by specifying some number $\varepsilon > 0$ and saying "*A non–negative number x is small iff $x < \varepsilon$*". To say that $\langle x_n \rangle_n$ is eventually small then means that from some point onwards all the x_n's are small, i.e

$$\exists N\ \forall n \geq N\ [x_n < \varepsilon]$$

This must be true no matter what gauge $\varepsilon > 0$ of "smallness" you use. Thus:

If $\langle x_n \rangle_n$ is a sequence of non–negative real numbers, we say

$$x_n \to 0 \qquad \Longleftrightarrow \qquad \forall \varepsilon > 0\ \exists N\ \forall n \geq N\ [x_n < \varepsilon]$$

We also write

$$\lim_{n \to \infty} x_n = 0$$

Thus $x_n \to 0$ iff given any $\varepsilon > 0$ it is possible to find a natural number N such that

$$x_n < \varepsilon \qquad \text{whenever } n \geq N$$

The number N typically depends on ε. The smaller $\varepsilon > 0$, the greater N usually has to be.

Example 2.2.5 We show that if $x_n := \frac{1}{\sqrt{n}}$, then $\lim_n x_n = 0$.
Let $\varepsilon > 0$. We must show that eventually $x_n < \varepsilon$, i.e. that there is $N \in \mathbb{N}$ such that

$$x_n < \varepsilon \qquad \text{whenever } n \geq N$$

Now note that, by the properties of an ordered field, $x_n < \varepsilon$ iff $n > \varepsilon^{-2}$. Proceed therefore as follows: By the Archimedean Property of the real numbers — which follows from the Completeness Axiom — there is an $N \in \mathbb{N}$ such that $N > \varepsilon^{-2}$. Note that this N depends on ε: the smaller ε, the greater N has to be, e.g. if $\varepsilon = \frac{1}{10}$, we can take $N = 101$ (or any larger integer), whereas if $\varepsilon = \frac{1}{100}$ we must take N to be at least 10 001. The point is that no matter how restrictive your definition of "small" is, there is a strategy for turning your ε into an N: Simply take N to be any integer $> \varepsilon^{-2}$.

Now if $n \geq N$, then the properties of an ordered field give

$$x_n \leq x_N < \varepsilon$$

and thus we have shown that for any ε there exists $N \in \mathbb{N}$ such that

$$x_n < \varepsilon \qquad \text{whenever } n \geq N$$

which is what is required.

□

Exercise 2.2.6 Let $p > 0$, and define $x_n := \frac{1}{n^p}$ for $n \in \mathbb{N}$. It is intuitively clear that $x_n \to 0$, but this needs proof... and here it is:

(a) Let $\varepsilon > 0$. Explain why there is an $N \in \mathbb{N}$ such that $N > \varepsilon^{-1/p}$.

(b) Now show that if $n \geq N$, $x_n < \varepsilon$.

(c) Explain why you have now shown that $\lim_n \frac{1}{n^p} = 0$.

Here, as in the previous example, we have presented an algorithm for determining an N from an ε: N is any integer $> \varepsilon^{-1/p}$.

□

2.2.3 Formal Definition of Convergence of Sequences

It is now rather simple to define convergence of arbitrary sequences in $\mathbb{R}$: To say that $x_n \to x$ means that the distance between x_n and x converges to 0, i.e.

$$x_n \to x \qquad \Leftrightarrow \qquad |x_n - x| \to 0$$

Now the distance $|x_n - x|$ between x_n and x is non–negative, so we already know what $|x_n - x| \to 0$ means from the previous subsection: It means $\forall \varepsilon > 0\ \exists N\ \forall n \geq N[|x_n - x| < \varepsilon]$. Thus

Definition 2.2.7 If $\langle x_n \rangle_n$ is a sequence of real numbers , we say

$$x_n \to x \qquad \Longleftrightarrow \qquad \forall \varepsilon > 0 \; \exists N \; \forall n \geq N \; [|x_n - x| < \varepsilon]$$

We then say that $\langle x_n \rangle_n$ is a *convergent* sequence, with *limit* x. We also write

$$x = \lim_n x_n \qquad \text{for} \qquad x_n \to x$$

A sequence which is not convergent is said to be *divergent.*

Thus $x_n \to x$ iff given any $\varepsilon > 0$ it is possible to find a natural number N such that

$$|x_n - x| < \varepsilon \qquad \text{whenever } n \geq N$$

The number N typically depends on ε. The smaller $\varepsilon > 0$, the greater N usually has to be.

To reiterate: $\langle x_n \rangle_n \to x$ iff for any $\varepsilon > 0$, all but finitely many terms lie within a distance of ε of x.

Remarks 2.2.8 Here is an alternative (topological) way of looking at the definition of convergence: Let U be an *open* interval containing x, so that x is not an endpoint of U. Then it is easy to see that there is a $\varepsilon > 0$ so that $(x - \varepsilon, x + \varepsilon) \subseteq U$. Now note that

$$|x_n - x| < \varepsilon \quad \Longleftrightarrow \quad x - \varepsilon < x_n < x + \varepsilon \quad \Longleftrightarrow \quad x_n \in (x - \varepsilon, x + \varepsilon) \subseteq U$$

We therefore see that

$x_n \to x$ if and only if for any open interval U containing x:

$x_n \in U$ eventually, i.e. at most finitely many x_n lie outside U.

This characterization of convergence mentions only open intervals, and suggests how one can define convergence in spaces more general than $\mathbb{R}$.

□

Again we note that the choice of N will generally depend on ε: The smaller ε is, the greater N must generally be. Another example will make this clear:

Example 2.2.9 Consider the sequence $\langle x_n \rangle$ given by $x_n = \frac{3n+1}{2n+3}$. By some algebraic manipulation, we see that

$$x_n = \frac{3n+1}{2n+3} = \frac{3 + \frac{1}{n}}{2 + \frac{3}{n}}$$

and since know $\frac{1}{n} \to 0$ as $n \to \infty$, we *guess* that $\lim_n x_n = \frac{3}{2}$. (Later, we will see that the above manipulations are perfectly acceptable. Right now, we don't know that yet, because we haven't proved it yet.)

But having guessed the answer, we can *prove* that it is correct, from the definition of convergence. So *prove* that $\lim_n x_n = \frac{3}{2}$. We must show that, given *any* $\varepsilon > 0$, we can find an $N \in \mathbb{N}$ such that $|x_n - \frac{3}{2}| < \varepsilon$ whenever $n \geq N$.

A little algebra shows that $|x_n - \frac{3}{2}| = \frac{7}{4n+6}$. Suppose that we are given $\varepsilon = \frac{1}{10}$. Then we must find an N such that $\frac{7}{4n+6} < \frac{1}{10}$ whenever $n \geq N$. Now

$$\frac{7}{4n+6} < \frac{1}{10} \qquad \text{iff} \qquad n > 16$$

Thus $|x_n - \frac{3}{2}| < \frac{1}{10}$ whenever $n \geq 17$. (If $n > 16$, then $n \geq 17$, since n is an integer.) So if we take $N = 17$, then $|x_n - \frac{3}{2}| < \frac{1}{10}$ for all $n \geq N$. Of course, we can take any larger N as well. If $N = 25$, then also $|x_n - \frac{3}{2}| < \frac{1}{10}$ whenever $n \geq N$ (because $n \geq 25$ implies that $n \geq 17$).

Does this prove that $\lim_n x_n = \frac{3}{2}$? No! We have merely shown that the requirements of the definition of limit can be fulfilled for the particular case $\varepsilon = \frac{1}{10}$. What we have to do is to show that the requirements can be fulfilled for *every* $\varepsilon > 0$.

Exercise: *Reasoning as above, show that if* $\varepsilon = \frac{1}{100}$*, then we can take* $N = 174$*, and if* $\varepsilon = \frac{1}{1000}$*, we can take* $N = 1749$*.*

We have now found N's that fulfil the requirements of the definition of limit for the cases $\varepsilon = \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$, respectively. But, to prove that $\lim_n x_n = \frac{3}{2}$, we must find such an N for every possible $\varepsilon > 0$. There are infinitely many such ε, so doing this case by case is impossible: We need a *general* argument.

Let $\varepsilon > 0$ be arbitrary. We want to show that we can find an N such that

$$|x_n - \tfrac{3}{2}| = \frac{7}{4n+6} < \varepsilon \qquad \text{for all } n \geq N$$

Clearly

$$\frac{7}{4n+6} < \varepsilon \qquad \text{iff} \qquad n > \frac{\frac{7}{\varepsilon} - 6}{4}$$

Choose N to be the least integer which is greater than $\frac{\frac{7}{\varepsilon}-6}{4}$. Such an N exists by the Archimedean Property of the reals. Then N fulfils the requirements of the definition of limit for ε. Note that N depends on ε.

Again, we have presented an *algorithm* (or recipe) for obtaining an N from an ε: Take N to be any integer that is greater than $\frac{\frac{7}{\varepsilon}-6}{4}$.

□

The following exercise will often be useful:

Exercise 2.2.10 Properties of absolute value:

(a) Triangle inequality: Show that $|x + y| \leq |x| + |y|$ for all $x, y \in \mathbb{R}$. (Hint: $|x| = \sqrt{x^2}$.)

(b) Show that $\big|\, |x| - |y| \,\big| \leq |x - y|$ for all $x, y \in \mathbb{R}$. (Hint: $|x| = |(x - y) + y|$. Now use the triangle inequality.)

□

Example 2.2.11 Suppose that we know that $x_n \to x$. We show that then also $x_n^2 \to x^2$. First note that the set $\{x_n : n \in \mathbb{N}\}$ is *bounded*: There is a number $K > 0$ such that $|x_n| \leq K$ for all $n \in \mathbb{N}$. For we may find N_1 such that whenever $n \geq N_1$, we have $|x_n - x| < 1$ (definition of convergence with $\varepsilon = 1$). Thus $|x_n| < |x| + 1$ for all $n \geq N_1$ (by Exercise 2.2.10(b))
Define

$$K = \max\{|x_1|, |x_2|, \ldots, |x_{N_1}|, |x| + 1\}$$

Note that if $n \in \mathbb{N}$, then $|x_n| \leq K$: For if $n \leq N_1$, then certainly $|x_n| \leq K$. On the other hand, if $n > N_1$, then $|x_n| < |x| + 1 \leq K$ as well. It follows that $\{x_n : n \in \mathbb{N}\}$ is bounded (by K).

We are now ready to show that $x_n^2 \to x^2$. Let $\varepsilon > 0$. Put $\bar{\varepsilon} = \frac{\varepsilon}{2K}$. Because $x_n \to x$, we can choose an N such that $n \geq N$ implies $|x_n - x| < \bar{\varepsilon}$. Then

$$|x_n^2 - x^2| = |x_n - x| \cdot |x_n + x| \leq |x_n - x| \cdot (|x_n| + |x|) \leq 2K|x_n - x| < \varepsilon$$

We have can thus find, for any $\varepsilon > 0$, an N such that: whenever $n \geq N$, we have $|x_n^2 - x^2| < \varepsilon$. And therefore we have shown that $x_n^2 \to x^2$ when $x_n \to x$.

The algorithm for determining N from ε works as follows: We are *given* that $x_n \to x$, and thus we have an algorithm, call it AL_1 for determining an N from an ε for the sequence $\langle x_n \rangle_n$. Above, we apply AL_1 to $\varepsilon = 1$ to obtain N_1. We then use N_1 to find K. And then we apply AL_1 once again to $\bar{\varepsilon} = \varepsilon/2K$ to obtain the required N.

□

Exercise 2.2.12 Comment on the following arguments:

(a) Argument 1:

> *Let $x_n = (-1)^n$, let $\varepsilon = 2$, and let $N = 1$. If $n \geq N$, then $|x_n - 0| = 1 < \varepsilon$. Hence $|x_n - 0| < \varepsilon$ for all $n \geq N$, and so* $\lim_n x_n = 0$

(b) Argument 2:

> *Let $x_n = (-1)^n$. We show that $\langle x_n \rangle_n$ does* not converge. *We do this by contradiction: For suppose that $x_n \to x$. Let $\varepsilon = \frac{1}{2}$. Then we can find N such that $n > N$ implies $|x_n - x| < \varepsilon$. This means that $|1 - x| < \frac{1}{2}$* and $|-1 - x| = |1 + x| < \frac{1}{2}$. *It follows that*
>
> $$2 = |1 - (-1)| \leq |1 - x| + |x - (-1)| < \frac{1}{2} + \frac{1}{2}$$
>
> *i.e. $2 < 1$, a contradiction.*
> *Hence there is no x such that $x_n \to x$.*

□

Example 2.2.13 Let $x_n = \frac{2n^3+5}{n^3-n+1}$. We want to show that $x_n \to 2$. So let $\varepsilon > 0$ be given. We want to find an N such that

$$\Big|\frac{2n^3 + 6}{n^3 - n + 1} - 2\Big| = \frac{2n + 4}{n^3 - n + 1} < \varepsilon$$

Solving for n in terms of ε will involve solving a cubic, which is tricky. However, we do not need to do this. Instead we obtain a simpler upper bound for $\frac{2n+4}{n^3-n+1}$, a bound of the form $\frac{Cn}{n^3} = \frac{C}{n^2}$, for some positive constant C. Note that $2n + 4 \leq 3n$ for all $n \geq 4$. Also note that $n^3 - n + 1 \geq \frac{1}{2}n^3$ for $n \geq 1$ (because $n^3 \geq 2n - 2$ for $n \geq 1$). Hence $\frac{2n+4}{n^3-n+1} \leq \frac{3n}{\frac{1}{2}n^3} = \frac{6}{n^2}$ for all $n \geq 4$.

It therefore suffices to find $N \geq 4$ so that $\frac{6}{n^2} \leq \varepsilon$ whenever $n \geq N$ (for if $\frac{6}{n^2} \leq \varepsilon$, then also $\frac{2n+4}{n^3-n+1} < \varepsilon$, provided $n \geq 4$.) Hence let N be an integer which is $> \max\{4, \sqrt{\frac{6}{\varepsilon}}\}$.

Now that we have figured out how to determine an N from given ε, we can now write down the

FORMAL PROOF: Let $\varepsilon > 0$. Choose an integer $N > \max\{4, \sqrt{\frac{6}{\varepsilon}}\}$. If $n \geq N$, then $n > \sqrt{\frac{6}{\varepsilon}}$, and so $\frac{6}{n^2} < \varepsilon$. Now since $n \geq 4$, we have $2n + 4 \leq 3n$. Also $n^3 - n + 1 \geq \frac{1}{2}n^3$ for $n \geq 1$, and hence $\frac{2n+4}{n^3-n+1} \leq \frac{3n}{\frac{1}{2}n^3} = \frac{6}{n^2} < \varepsilon$ when $n \geq N$. It follows that

$$\Big|\frac{2n^3+6}{n^3-n+1} - 2\Big| = \frac{2n+4}{n^3-n+1} < \varepsilon \qquad \text{whenever } n \geq N$$

Hence $\lim_n \frac{2n^3+6}{n^3-n+1} = 2$.

□

Exercises 2.2.14 1. Find an N fulfilling the conditions for the given convergent sequence $\langle x_n \rangle$ and the given ε.

(a) $x_n = \frac{\sin 2n}{n}$, $\varepsilon = 0.1$.

(b) $x_n = \frac{2^n}{n!}$, $\varepsilon = 0.001$. (First prove that x_n is a *decreasing* sequence for $n \geq 1$ — i.e. that $x_{n+1} \leq x_n$ for all $n \geq 1$ — and then find N by "brute force".)

(c) $x_n = \frac{4n^3+3n}{n^3-6}$, $\varepsilon = 10^{-6}$.

□

Exercises 2.2.15 (1.) Show that if $\langle x_n \rangle_n = \langle c, c, c, \ldots \rangle$ is a constant sequence in $\mathbb{R}$, then $x_n \to c$.

(2.) Show that if $x_n \to x$ in $\mathbb{R}$, then $|x_n| \to |x|$.

(3.) Consider the following sequence $\langle x_n \rangle_n$:

$$0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, \ldots, 1\underbrace{0, 0, \ldots, 0}_{n \text{ zeroes}}, 1, \underbrace{0, 0, \ldots, 0}_{n+1 \text{ zeroes}}, 1, \ldots$$

Does this sequence converge? Carefully explain your answer.

□

We also sometimes say that a sequence converges to $\pm\infty$.

To say that $x_n \to \infty$ means $\langle x_n \rangle_n$ is "large" eventually.

The notion "large" is subjective, so we will demand that it holds for absolutely anybody's idea of "large". Specifically, suppose you define "large" by specifying some number $K > 0$ and saying "*A number x is large iff $x > K$*". To say that $\langle x_n \rangle_n$ is eventually large then means that from some point onwards all the x_n's are large, i.e

$$\exists N\ \forall n \geq N\ [x_n > K]$$

This must be true no matter what gauge $K > 0$ of "largeness" you use. Thus:

Definition 2.2.16 If $\langle x_n \rangle_n$ is a sequence of real numbers, we say

$$x_n \to \infty \qquad \Longleftrightarrow \qquad \forall K > 0 \; \exists N \; \forall n \geq N \; [x_n > K]$$

We say that $x_n \to -\infty$ iff $-x_n \to +\infty$.

Remarks 2.2.17 (a) Note that a sequence which "converges to ∞" is nevertheless a *divergent* sequence: $\lim_n x_n$ does *not* exist (as a real number). Nevertheless, this divergence seems not to be too badly behaved. We therefore say that $\lim_n x_n$ exists in the *extended sense*.

(b) We may also write $\lim_n x_n = \infty$ instead of $x_n \to \infty$, etc.

(c) Do NOT confuse ∞ with a real number. ∞ is meaningless by itself, and we haven't "created" or discovered a new number (as arguably, we *have* done in the case of $i = \sqrt{-1}$). To say $\lim_n x_n = \infty$ is simply a short hand for "Given any number $K > 0$, eventually $x_n > K$". Note that infinity isn't mentioned in the definition of $x_n \to \infty$.

(d) In particular, note that the limit theorems obtained in this and the previous section do not apply to ∞.

□

Example 2.2.18 Let $x_n = \sqrt{n}$. We want to show that $x_n \to \infty$. So let $K > 0$ be given. We must find an N such that $x_n > K$ whenever $n > N$. It is easy to see that $N = K^2$ will do the trick.

□

Exercise 2.2.19 Suppose that $\langle x_n \rangle_n$ is a sequence of *strictly positive* reals. Prove that $\lim_n x_n = \infty$ iff $\lim_n \frac{1}{x_n} = 0$. Why do we require that the x_n be strictly positive, rather than just non–zero?
[Hint: ($\Rightarrow$) Given $\varepsilon > 0$, choose N such that $x_n > \frac{1}{\varepsilon}$ whenever $n \geq N$.]

□

We wrap up this section with two "obvious" facts:

Proposition 2.2.20 *A sequence can have at most one limit.*

Proof: Suppose that $x_n \to x$ and that $x_n \to y$, where $x \neq y$. Let $0 < \varepsilon < \frac{|x-y|}{2}$. First choose N_x such that $|x_n - x| < \varepsilon$ whenever $n \geq N_x$. Then choose N_y such that $|x_n - y| < \varepsilon$ whenever $n \geq N_y$. Now let $N = \max\{N_x, N_y\}$. Thus:

$$\text{If } n \geq N, \text{ then both} \quad |x_n - x| < \varepsilon \quad \text{and} \quad |x_n - y| < \varepsilon$$

Hence

$$|x - y| \leq |x - x_N| + |x_N - y| < 2\varepsilon < |x - y|$$

Thus, assuming that a sequence has two distinct limits x, y, we have concluded that $|x - y| < |x - y|$, a contradiction.

⊣

Definition 2.2.21 A subset $A \subset \mathbb{R}$ is bounded if it is contained in a finite interval. Equivalently, A is bounded if and only if there is a number K such that $|a| \leq K$ for all $a \in A$ (in which case $A \subseteq [-K, K]$).

Note that Example 2.2.11 contains a useful proposition:

Proposition 2.2.22 *Any convergent sequence is bounded.*

Proof: Suppose that $x_n \to x$. Choose N_1 such that whenever $n \geq N_1$, we have $|x_n - x| < 1$ (definition of convergence with $\varepsilon = 1$). Thus $|x_n| < |x| + 1$ for all $n \geq N_1$ (by Exercise 2.2.10(b))
Now define

$$K = \max\{|x_1|, |x_2|, \ldots, |x_{N_1}|, |x| + 1\}$$

Note that if $n \in \mathbb{N}$, then $|x_n| \leq K$: For if $n \leq N_1$, then certainly $|x_n| \leq K$. On the other hand, if $n > N_1$, then $|x_n| < |x| + 1 \leq K$ as well. It follows that $\{x_n : n \in \mathbb{N}\}$ is bounded by K.

⊣

Exercise 2.2.23 Here are some variations on the definition of limit: These are *not serious* definitions, but merely intended to give you a feel for the role of the quantifiers in the definition of limit.

(i) Define a sequence $\langle x_n \rangle_n$ of real numbers to be *W–convergent* to x if and only $\forall N \in \mathbb{N} \exists \varepsilon > 0 \forall n \geq N[|x_n - x| < \varepsilon]$.

(ii) Define a sequence $\langle x_n \rangle_n$ of real numbers to be *S–convergent* to x if and only $\exists N \in \mathbb{N} \forall \varepsilon > 0 \forall n \geq N[|x_n - x| < \varepsilon]$.

Now do the following:

(a) Describe all the W–convergent sequences, and show that every convergent sequence is W–convergent. Give an example of a W–convergent sequence which is not S–convergent.

(b) Describe all the S–convergent sequences, and show that every S–convergent sequence is convergent. Give an example of an S–convergent sequence which is not convergent.

(c) Show that W–limits may not be unique, but that S–limits are unique.

□

2.3 Arithmetic, Order and Convergence

2.3.1 Arithmetic and Convergence

It can be quite difficult to prove that a sequence x_n converges to a particular limit x if we use only the definition of convergence (i.e. Definition 2.2.7). For example, to show that $\frac{2n^3+5}{n^3-n+1} \to 2$ directly from the definition was quite tricky (cf. Example 2.2.13).

Yet it is easy to "see" that $\frac{2n^3+5}{n^3-n+1} \to 2$:

$$\frac{2n^3 + 5}{n^3 - n + 1} = \frac{2 + \frac{5}{n^3}}{1 - \frac{1}{n^2} + \frac{1}{n^3}}$$

Now $\frac{1}{n^3} \to 0$, so the numerator $(2 + \frac{5}{n^3}) \to 1$. Similarly, $1 - \frac{1}{n^2} + \frac{1}{n^3} \to 1$. Thus $\frac{2n^3+5}{n^3-n+1} \to \frac{2}{1}$.

If we analyze the above "proof", we see that we have made the following assumptions:

(i) If $x_n \to x$ and $y_n \to y$, then $(x_n+y_n) \to (x+y)$. Thus $\lim_n(x_n+y_n) = \lim_n x_n + \lim_n y_n$, i.e.

The limit of a sum is the sum of the limits

It doesn't matter if we first add the x's to the y's and then take the limit, or if we first take the limits of the x's and the y's, and then add those. Addition *commutes* with limit.

(ii) If $x_n \to x$, and $y_n \to y$, then $\frac{x_n}{y_n} \to \frac{x}{y}$ (assuming $y \neq 0$). Thus $\lim_n \frac{x_n}{y_n} = \frac{\lim_n x_n}{\lim_n y_n}$, and division commutes with limit.

Of course, this needs proof.

Theorem 2.3.1 *Let $\langle x_n\rangle_n, \langle y_n\rangle_n$ be sequences in $\mathbb{R}$, with $x_n \to x$, $y_n \to y$. Also let $\alpha \in \mathbb{R}$. Then*

(a) $(x_n + y_n) \to x + y$;

(b) $\alpha x_n \to \alpha x$;

(c) $x_n y_n \to xy$;

(d) $\frac{1}{x_n} \to \frac{1}{x}$ *(provided $x_n \neq 0$ for $n \in \mathbb{N}$, and $x \neq 0$);*

(e) $\frac{x_n}{y_n} \to \frac{x}{y}$ *(provided $y_n \neq 0$ for $n \in \mathbb{N}$, and $y \neq 0$).*

To help you understand what exactly must be accomplished, we'll leave the first two as exercises. Do this exercise NOW!

Exercise 2.3.2 (a) Suppose that $x_n \to x$, $y_n \to y$ in $\mathbb{R}$. We will show that $(x_n+y_n) \to x+y$.

(i) We must show that, given any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that $|(x_n+y_n)-(x+y)| < \varepsilon$ whenever $n \geq N$. Make sure you understand this.

(ii) Explain why we can find N_1 such that $|x_n - x| < \frac{\varepsilon}{2}$ whenever $n \geq N_1$.

(iii) Similarly, we can find an N_2 such that the same holds for y_n.

(iv) Let $N = \max\{N_1, N_2\}$. Use the triangle inequality to deduce that $|(x_n + y_n) - (x + y)| < \varepsilon$ when $n \geq N$.

(v) Conclude that $\lim_n(x_n + y_n) = x + y$, as required.

(b) Let $x_n \to x$ in $\mathbb{R}$, and suppose that $\alpha \in \mathbb{R}$. Show that $\alpha x_n \to \alpha x$.
[Hint: If $\alpha \neq 0$, choose $N \in \mathbb{N}$ such that $|x_n - x| < \frac{\varepsilon}{|\alpha|}$ whenever $n > N$. Also, don't forget the case $\alpha = 0$.]

□

Proof of Theorem 2.3.1: I hope you did Exercise 2.3.2...
We need only prove (c)–(e), as (a), (b) were dealt with there. So suppose that $x_n \to x, y_n \to y$ in $\mathbb{R}^1$, and let $\varepsilon > 0$. Note that

$$x_n y_n - xy = x_n(y_n - y) + y(x_n - x)$$

Now because the sequence $\langle x_n\rangle_n$ converges, it is bounded (by Proposition 2.2.22), and thus there is a $K_1 > 0$ such that $|x_n| \leq K_1$, for all $n \in \mathbb{N}$, and also $|x| \leq K_1$. Similarly, there is $K_2 > 0$ such that $|y_n| \leq K_2$ for all $n \in \mathbb{N}$, and also $|y| \leq K_2$.

Now choose N_1, N_2 such that

$$n > N_1 \Rightarrow |x_n - x| \leq \frac{\varepsilon}{2K_2}$$
$$n > N_2 \Rightarrow |y_n - y| \leq \frac{\varepsilon}{2K_1}$$

Put $N = \max\{N_1, N_2\}$. If $n > N$, we have

$$|x_n(y_n - y)| = |x_n| \cdot |y_n - y| < K_1 \cdot \frac{\varepsilon}{2K_1} = \frac{\varepsilon}{2}$$

Similarly, $n > N$ implies $|y(x_n - x)| < \frac{\varepsilon}{2}$. Thus

$$n > N \Rightarrow |x_n y_n - xy| \leq |x_n(y_n - y)| + |y(x_n - x)| < \varepsilon$$

and this proves (c).

To prove (d), suppose that $\langle x_n\rangle_n$ is a sequence of non–zero real numbers which converges to x, where also $x \neq 0$. We first show that the sequence $\langle \frac{1}{x_n}\rangle_n$ is bounded, i.e. that there is a K such that, for all $n \in \mathbb{N}$, we have $|\frac{1}{x_n}| \leq K$. For, since $|x| \neq 0$, we can choose $K_1 > 0$ such that $\frac{1}{K_1} < |x|$ (by the Archimedean Property). Now since $x_n \to x$, we also have $|x_n| \to |x|$ (by Exercise 2.2.15), and thus $|x_n| > \frac{1}{K_1}$ eventually. Formally, there is N such that $n > N$ implies $\frac{1}{K_1} < |x_n|$. Now define

$$K = \max\{\frac{1}{|x_1|}, \ldots, \frac{1}{|x_N|}, K_1\}$$

Then $\frac{1}{|x_n|} \leq K$ for *all* $n \in \mathbb{N}$, proving that $\langle \frac{1}{x_n}\rangle_n$ is bounded (by K).

Next, note that

$$\Big|\frac{1}{x_n} - \frac{1}{x}\Big| = \frac{|x - x_n|}{|x_n x|}$$

Given now an $\varepsilon > 0$, we can find an N such that

$$n > N \Rightarrow |x_n - x| < \frac{\varepsilon}{K^2}$$

because $x_n \to x$. It follows that

$$n > N \Rightarrow \Big|\frac{1}{x_n} - \frac{1}{x}\Big| = \frac{|x - x_n|}{|x_n x|} < K^2 \cdot \frac{\varepsilon}{K^2} = \varepsilon$$

as $\frac{1}{|x_n x|} \leq K^2$. Thus $\frac{1}{x_n} \to \frac{1}{x}$, as required. This proves (d).

Finally, (e) is an easy consequence of (c) and (d): Just notice that $\frac{x_n}{y_n} = x_n \cdot \frac{1}{y_n}$.

⊣

Exercise 2.3.3 Suppose that $\langle y_n\rangle_n$ is a *non–negative* sequence in $\mathbb{R}$ which converges to y. We show that then also $\sqrt{y_n} \to \sqrt{y}$.

(a) We consider, separately, two cases: $y > 0$ and $y = 0$. First assume that $y = 0$, i.e. that $y_n \to 0$. Prove that in that case also $\sqrt{y_n} \to 0$.

(b) Next assume that $y > 0$. Explain why there is a $K > 0$ such that $y \geq K$ and such that $y_n \geq K$ *eventually.*

(c) Now notice that $(\sqrt{y_n} - \sqrt{y})(\sqrt{y_n} + \sqrt{y}) = (y_n - y)$, and that $(\sqrt{y_n} + \sqrt{y}) \geq 2\sqrt{K}$. Use these facts to show that $\sqrt{y_n} \to \sqrt{y}$ in this case also.
[Hint: Choose N such that $|y_n - y| < 2\sqrt{K}\varepsilon$ whenever $n \geq N$.]

□

Exercise 2.3.4 Let $\langle x_n \rangle_n$ be an arbitrary sequence of real numbers. Construct a new sequence $\langle y_n \rangle_n$ of *Cesaro means* by defining

$$y_n := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

(a) Show that if $\langle x_n \rangle_n$ converges, then so does $\langle y_n \rangle_n$, and to the same limit.
[Hint: Suppose $x_n \to x$. First choose M such that $|x_n - x| < \frac{\varepsilon}{2}$ whenever $n \geq M$. Then choose $N \geq M$ such that $\frac{|x_k - x|}{N} < \frac{\varepsilon}{2M}$ for all $k < M$. Then show

$$|y_n - x| \leq \left(\frac{|x_1 - x|}{n} + \cdots + \frac{|x_M - x|}{n} \right) + \left(\frac{|x_{M+1} - x|}{n} + \ldots \frac{|x_n - x|}{n} \right)$$

and let $n \geq N$.]

(b) *Prove or Disprove:* If $\langle y_n \rangle_n$ converges, then so does $\langle x_n \rangle_n$, and to the same limit.

2.3.2 Order, Completeness and Convergence

Note that

$$x_n \to x \qquad \text{iff} \qquad \forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n \geq N [x - \varepsilon < x_n < x + \varepsilon]$$

This is because $|y - x| < \varepsilon$ iff $x - \varepsilon < y < x + \varepsilon$. Thus to say that $x_n \to x$ means that, for any $\varepsilon > 0$, eventually the sequence $\langle x_n \rangle$ lies in the open interval $(x - \varepsilon, x + \varepsilon)$.

The next theorem is left as an exercise:

Theorem 2.3.5 (Sandwich Theorem, or Squeeze Theorem)
Suppose that $\langle x_n \rangle_n, \langle y_n \rangle_n$ and $\langle z_n \rangle_n$ are sequences in $\mathbb{R}$ which satisfy the following conditions:

(i) $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$ (or merely eventually*);*

(ii) There is $l \in \mathbb{R}$ such that $x_n \to l$ and $z_n \to l$.

Then also $y_n \to l$.

Exercise 2.3.6 The aim of this exercise is to prove the Sandwich Theorem. So let $\varepsilon > 0$. We must show that there is $N \in \mathbb{N}$ such that $|y_n - l| < \varepsilon$ whenever $n > N$, or equivalently, that $l - \varepsilon < y_n < l + \varepsilon$ whenever $n > N$.

(a) Assume first that $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$. Explain why there is $N_1 \in \mathbb{N}$ such that whenever $n > N_1$, we have $l - \varepsilon < x_n < l + \varepsilon$.

(b) Now explain why there is an $N \in \mathbb{N}$ such that whenever $n > N$, we have *both* $l - \varepsilon < x_n < l + \varepsilon$ *and* $l - \varepsilon < z_n < l + \varepsilon$.

(c) Now explain why also $l - \varepsilon < y_n < l + \varepsilon$ whenever $n > N$.

(d) The Theorem has now been proved for the case where $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$. Modify your proof slightly to show that the Theorem remains true if we have $x_n \leq y_n \leq z_n$ *eventually*.

□

Example 2.3.7 We use the Sandwich Theorem to show that if $|x| < 1$, then $x^n \to 0$ (as $n \to \infty$). This is obvious if $x = 0$. Now if $0 < |x| < 1$, then $\frac{1}{|x|} > 1$, i.e $\frac{1}{|x|} = 1 + h$ for some positive h. Thus $|x|^n = \frac{1}{(1+h)^n}$. Now by the Binomial Theorem,

$$(1+h)^n = \binom{n}{0}h^0 + \binom{n}{1}h + \binom{n}{2}h^2 + \cdots + \binom{n}{n}h^n \geq nh$$

Hence $0 \leq |x|^n \leq \frac{1}{nh}$. Now we proved earlier that $\frac{1}{n} \to 0$, and thus $\frac{1}{nh} \to 0$ as well. Put $a_n = 0$, $b_n = \frac{1}{nh}$ for all n. Then

$$a_n \to 0, \quad b_n \to 0 \quad \text{and} \quad a_n \leq |x|^n \leq b_n$$

By the Sandwich Theorem, also $|x|^n \to 0$. It follows easily that $x^n \to 0$ as well.

□

Exercise 2.3.8 Show that $\sqrt[n]{n} \to 1$ as $n \to \infty$.
[Hint: Let $x_n = \sqrt[n]{n} - 1$, and use the binomial theorem to prove that $n \geq \frac{n(n-1)}{2}x_n^2$. Then use the Sandwich Theorem to show that $x_n \to 0$.]

□

Next, we consider *monotone* sequences:

Definition 2.3.9 Let $\langle x_n \rangle_n$ be a sequence in $\mathbb{R}$.

(a) $\langle x_n \rangle_n$ is said to be an *increasing sequence* if and only if $n \leq m$ implies $x_n \leq x_m$, i.e. iff

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n \leq \ldots$$

(b) $\langle x_n \rangle_n$ is said to be a *strictly increasing sequence* if and only if $n < m$ implies $x_n < x_m$, i.e. iff

$$x_1 < x_2 < x_3 < \cdots < x_n < \ldots$$

(c) *Decreasing* and *strictly decreasing* sequences are defined by replacing $\leq$ with $\geq$, and $<$ with $>$.

(d) A *monotone sequence* is one that is either increasing or decreasing. A *strictly monotone sequence* is one that is either strictly increasing or strictly decreasing.

Warning: In the literature, what we call an increasing sequence is often called a *non–decreasing* sequence, and what we call a strictly increasing sequence is often called an *increasing* sequence.

Examples 2.3.10 (a) $1, 2, 3, \ldots$, is strictly increasing, and thus increasing.

(b) Any constant sequence is both increasing and decreasing. However, it is neither strictly increasing, nor strictly decreasing.

(c) The sequence $1, 2, 1, 2, 1, 2, \ldots$ is neither increasing nor decreasing.

□

The following fundamental — and surprisingly useful — fact is just the Completeness Axiom, couched in sequence terminology.

Theorem 2.3.11 *Let $\langle x_n \rangle_n$ be a monotone sequence in $\mathbb{R}$. Then $\langle x_n \rangle_n$ converges if and only if it is bounded.*
Moreover, if $\langle x_n \rangle_n$ is a bounded increasing sequence, it converges to $\sup_n x_n$, whereas if $\langle x_n \rangle_n$ is a bounded decreasing sequence, it converges to $\inf_n x_n$.

Proof: ($\Rightarrow$) If $\langle x_n \rangle_n$ is any convergent sequence (monotone or not), then it is bounded, by Proposition 2.2.22.
($\Leftarrow$) Suppose that $\langle x_n \rangle_n$ is an increasing bounded sequence. By the Completeness Axiom, there is a number

$$x^* = \sup_n \{x_n : n \in \mathbb{N}\}$$

We will show that $x_n \to x^*$. So let $\varepsilon > 0$. Then $x^* - \varepsilon$ is *not* an upper bound of the set $\{x_n : n \in \mathbb{N}\}$ (because $x^* - \varepsilon$ is strictly less than the *least* upper bound x^*). It follows that there is $n_0 \in \mathbb{N}$ such that $x_{n_0} > x^* - \varepsilon$

Now if $n \geq n_0$ we have

(i) $x_n \geq x_{n_0}$, because $\langle x_n \rangle_n$ is increasing.

(ii) $x_n \leq x^*$, because x^* is an upper bound of the sequence elements. Hence $|x_n - x^*| = x^* - x_n \leq x^* - x_{n_0} < \varepsilon$.

Since we can find such an n_0 for every $\varepsilon > 0$, we have shown that

$$\forall \varepsilon > 0 \; \exists n_0 \in \mathbb{N} \; \forall n > n_0 [|x_n - x^*| < \varepsilon]$$

which is the same as saying $x_n \to x^*$.

We have therefore shown that any bounded increasing sequence converges. Suppose now that $\langle x_n \rangle_n$ is a bounded decreasing sequence. Then $\langle -x_n \rangle_n$ is a bounded increasing sequence, and thus converges. It is now easy to see that $\langle x_n \rangle_n$ converges as well.

⊣

Remarks 2.3.12 If $\langle x_n \rangle_n$ is an increasing sequence which converges to x, we often write

$$x_n \uparrow x \qquad \text{or} \qquad x = \uparrow \lim_n x_n$$

instead of $x_n \to x$. Similarly, if $\langle x_n \rangle_n$ is a decreasing convergent sequence, we write

$$x_n \downarrow x \qquad \text{or} \qquad x = \downarrow \lim_n x_n$$

□

Note that the Completeness Axiom, via preceding theorem, guarantees that a bounded monotone sequence has a limit, *even* if we have no way of directly determining what that limit is. This is the case in the following example:

Example 2.3.13 Define $x_n := (1 + \frac{1}{n})^n$. We show that $\langle x_n \rangle_n$ converges. By the preceding theorem, it suffices to show (i) that $\langle x_n \rangle_n$ is increasing, and (ii) that it is bounded. Now by the Binomial Theorem

$$\begin{aligned} x_n &= 1 + \frac{n}{1}\frac{1}{n} + \frac{n(n-1)}{2!}\frac{1}{n^2} + \cdots + \frac{n(n-1)\ldots 2.1}{n!}\frac{1}{n^n} \\ &= 1 + 1 + \frac{1}{2!}\Big(1 - \frac{1}{n}\Big) + \cdots + \frac{1}{n!}\Big(1 - \frac{1}{n}\Big)\Big(1 - \frac{2}{n}\Big)\ldots\Big(1 - \frac{n-1}{n}\Big) \end{aligned}$$

Similarly,

$$x_{n+1} = 1 + 1 + \frac{1}{2!}\Big(1 - \frac{1}{n+1}\Big) + \cdots + \frac{1}{(n+1)!}\Big(1 - \frac{1}{n+1}\Big)\Big(1 - \frac{2}{n+1}\Big)\ldots\Big(1 - \frac{n}{n+1}\Big)$$

Now we compare terms. x_n has $n+1$ terms, whereas, x_{n+1} has $n+2$ terms, all non–negative. x_n and x_{n+1} agree on the first two terms. Now if $2 < k \leq n+1$, then the k^{th} term of x_n is

$$\frac{1}{(k-1)!}\Big(1 - \frac{1}{n}\Big)\Big(1 - \frac{2}{n}\Big)\ldots\Big(1 - \frac{k-2}{n}\Big)$$

whereas the k^{th} term of x_{n+1} is

$$\frac{1}{(k-1)!}\Big(1 - \frac{1}{n+1}\Big)\Big(1 - \frac{2}{n+1}\Big)\ldots\Big(1 - \frac{k-2}{n+1}\Big)$$

It is therefore clear that the k^{th} term of x_n is less than the k^{th} term of x_{n+1}. Moreover, x_{n+1} has one more term, which is strictly positive. It follows that $x_n < x_{n+1}$.
Thus $\langle x_n \rangle_n$ is an increasing sequence.

Next, we show that $\langle x_n \rangle_n$ is bounded. Look again at the k^{th} term of x_n: We have

$$\frac{1}{(k-1)!}\Big(1 - \frac{1}{n}\Big)\Big(1 - \frac{2}{n}\Big)\ldots\Big(1 - \frac{k-2}{n}\Big) \leq \frac{1}{2^{k-2}}$$

This is (i), because

$$2^{k-2} = 2 \cdot 2 \cdot \cdots \cdot 2 \leq 2 \cdot 3 \cdot \cdots \cdot (k-1)$$

and (ii), because

$$\Big(1 - \frac{1}{n}\Big)\Big(1 - \frac{2}{n}\Big)\ldots\Big(1 - \frac{k-2}{n}\Big) \leq 1 \cdot 1 \cdot \cdots \cdot 1$$

It follows that

$$x_n \leq 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \leq 3$$

and thus that $x_n \leq 3$ for all n. Hence $\langle x_n \rangle_n$ is bounded.

We can now conclude that $\langle x_n \rangle_n$ converges, though we do not yet know precisely where it converges to.

If you stick the x_n into a calculator, you will see that $x_n \to e$, where $e = 2.7182818\ldots$ is the base of the natural logarithm.

□

Exercise 2.3.14 Consider the following inductively presented sequence

$$x_1 = 1 \qquad x_{n+1} = \sqrt{x_n + 1}$$

(a) Show that $x_n \leq 2$ for all $n \in \mathbb{N}$.

(b) Show that $\langle x_n \rangle$ is increasing.

(c) Conclude that $x = \lim_n x_n$ exists.

(d) Prove that $\lim_n \sqrt{x_n + 1} = \sqrt{x + 1}$.

(e) Conclude that $\lim_n x_n = \frac{1+\sqrt{5}}{2}$, the *Golden Ratio*.

[Hints: (a) Use mathematical induction. (b) Induction again: Assuming $x_n \leq x_{n+1}$, show that $x_{n+1} \leq x_{n+2}$. (d) Exercise 2.3.3]

□

Remarks 2.3.15 In the previous exercise, we were given an inductively defined sequence in the form

$$x_1 = c \qquad x_{n+1} = f(x_n)$$

where f is a *continuous* function. In that case, *if* the limit $x = \lim_n x_n$ exists, *then* it is a fixed point of f: it satisfies the equation

$$f(x) = x$$

This is because

$$x = \lim_n x_{n+1} = \lim_n f(x_n) = f(\lim_n x_n) = f(x)$$

Interchanging the order of function f and limit is permitted because f is continuous — something which we will discuss in more detail in a later chapter.

Beware, however: Blindly applying the above reasoning to the sequence

$$x_1 = 1 \qquad x_{n+1} = -x_n$$

yields $-x = x$ (where $x = \lim_n x_n$), and so $x = 0$. But, of course, the sequence $\langle x_n \rangle$ is just $1, -1, 1, -1, \ldots$, which does not converge. Therefore, before you can apply the above method to find the limit, you must be sure that the given sequence actually *has* a limit.

□

Exercise 2.3.16 Show that the following sequences converge, and hence find their limits:

(a) $x_1 = 1$, $x_{n+1} = \frac{n}{n+1} x_n^2$

(b) $x_1 = 1$, $x_{n+1} = (x_n + 1)/3$.

□

2.4 Representation of Real Numbers by Decimals

Exercise 2.4.1 We are very used to representing a real number by a decimal expansion, e.g. $\pi = 3.14159265\ldots$. When we investigated the structure of the reals, however, we decided that the reals *are* a complete ordered field — decimals weren't mentioned once. Right now, in our development of the reals, we do not yet know what we mean by an expression like $3.14159265\ldots$. So here are two questions:

(I) What, exactly, do we *mean* by an expression $3.14159265\ldots$? There are infinitely many numbers in this expansion. — Can we give this expression a precise meaning?

(II) Can every real number be represented by a decimal expansion?

We restrict ourselves to numbers in the unit interval $[0, 1)$. Recall that $[x]$ is defined to be the greatest integer which is $\leq x$, e.g. $[\pi] = 3$, $[\sqrt{2}] = [1.9] = [1] = 1$.

(a) Consider the expression $0.a_1a_2a_3a_4\ldots$, where the a_n are integers with $0 \leq a_n \leq 9$. We'd like to assign a meaning to this expression. Define $s_n = \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_n}{10^n}$. Show that the sequence $\langle s_n \rangle_n$ converges.

(b) Spend some time convincing yourself that it would be a good idea to *define*

$$0.a_1a_2a_3a_4\cdots = \lim_n s_n$$

(or else, if you disagree, find a different method which assigns a meaning to the expression $0.a_1a_2a_3a_4\ldots$).

(c) Moving on to Question (II), we want to show that every real number has a decimal expansion. So let $s \in [0, 1)$. Define two sequences $\langle x_n \rangle_n$ and $\langle a_n \rangle_n$ inductively, as follows: Put $x_1 = s$. Assuming that $x_1, \ldots, x_n$ and $a_1, \ldots, a_{n-1}$ have been defined, put

$$a_n = [10x_n] \qquad x_{n+1} = 10x_n - a_n$$

Find the first 5 terms of each sequence if $s = \frac{1}{7}$.

(d) Explain why $0 \leq x_n < 1$ for each n.

(e) Hence show that each a_n is an integer, and that $0 \leq a_n \leq 9$.

(f) Show (by induction) that $x_{n+1} = 10^n s - (10^{n-1}a_1 + 10^{n-2}a_2 + \cdots + a_n)$.

(g) As above, define $s_n = \sum_{k=1}^{n} \frac{a_k}{10^k}$. Conclude that $0 \leq s - s_n < 10^{-n}$.

(h) Hence show that $s_n \to s$, and thus that $s = 0.a_1a_2a_3a_4\ldots$.

□

So we have now shown that every real number has a decimal representation. Of course, we have not shown that such a representation is unique — Indeed it isn't. For example, $0.25 = 0.24999\ldots$. But it is only terminating[1] decimals that have another representation. If you think carefully, you will conclude that all reals have a unique non–terminating decimal representation.

[1] i.e. decimals that end, eventually, in all zeroes.

2.5 Introduction to Series

In this section, we are concerned with numerical expressions of the form

$$x_1 + x_2 + x_3 + \cdots + x_n + \ldots$$

which we may also write as

$$\sum_{k=1}^{\infty} x_k$$

Such an expression is called an *infinite series*, or just a *series*.

2.5.1 The Paradoxes of Zeno

The Greek philosopher Zeno (of Elea, ca. 450 BC) is responsible for several delightful paradoxes. The most well–known paradox involves the swift–footed warrior Achilles, whose rage is the cental motive of Homer's epic *The Iliad*, and a Tortoise who studied under Socrates. Briefly, the story has the following plot:

> The Tortoise challenges Achilles to a race, claiming that he would win over any distance, provided that Achilles give him a small head start. Achilles agrees, with alacrity, and a date is set. Just before the race begins, Achilles and the Tortoise engage in some idle chit–chat:
> TORTOISE: *You may as well concede the race now, oh Achilles. It is* logically *impossible for you to win.*
> ACHILLES: . . . (Though the son of a Goddess, Achilles doubts whether the Gods themselves are immune from the constraints of logic. . .)
> TORTOISE: *Let me demonstrate. You agree that you will run the the distance of my head start in quite a short time?*
>
> ACHILLES: *Indeed, a* very *short time. After I have run to your starting point, I will be practically upon you — you will be ahead only very slightly.*
> TORTOISE: *But you agree that I will be ahead. . . Nevertheless, you will cover the remaining distance between us quite quickly?*
> ACHILLES: *Indeed,* extremely quickly*. When I cover that distance you'll have inched forward hardly any distance at all.*
> TORTOISE:*But I will have inched forward some distance, and will therefore be ever so slightly ahead. . . Now you will presumably cover the new distance between us in very little time?*
> ACHILLES:*As you say in hardly any time at all. You will barely have moved. . . Oh. . .*
> TORTOISE: *I see you have spotted the problem, Achilles. No matter how fast you move, when you reach the spot where you saw me last, I will have moved ahead, be it ever so slightly. Thus you will never catch up with me.*
> ACHILLES: *It is as you say, oh Tortoise. . .*

says Achilles, and proceeds to win the race with ease. [2]

Let us examine the Tortoise's argument. We know, from experience, that the Tortoise is wrong. We also know, from experience, that logic never lets us down, only the faulty application of it. Thus we (rightly) suspect an error in the Tortoise's logic.

Suppose that Achilles and the Tortoise agree on a 50 meter race, and the Tortoise is to get a 9 meter head start. Suppose further that Achilles, in full battle gear, is able to attain a speed of 10m/s, and that the Tortoise can crawl at a very respectable 1m/s.

- Firstly, let's do a quick calculation to determine who wins. Achilles reaches the finish line in 5 seconds (50 meters at 10m/s). At that time the Tortoise, travelling at 1m/s, will only be at the 14 meter mark (9m + 5m). Thus Achilles wins, by 36m.

- Next, let us use a simple *algebraic* argument to determine at what time (call it T) Achilles overtakes the Tortoise. At time T, Achilles will have travelled a distance of $10T$, whereas the Tortoise will have crawled to the point $9 + T$. Thus $10T = 9 + T$, which implies that $T = 1$. Thus, Achilles overtakes the Tortoise at $T = 1$ seconds.

- Next, let us carefully examine the Tortoise's argument: Achilles will reach the Tortoise's starting point (the 9m mark) in $\frac{9}{10}$ of a second. At that stage the Tortoise will have crawled $\frac{9}{10}$ meters ahead. Achilles will cover this distance in a mere $\frac{9}{100}$ of a second, but the Tortoise will have used this time to inch forward another $\frac{9}{100}$ meters. Achilles will cover *this* distance in just $\frac{9}{1000}$ of a second, giving the Tortoise enough time to gain an additional $\frac{9}{1000}$ meters.... The Tortoise's argument can now be summed up as follows: From the above, we can see that Achilles will catch up with the Tortoise at time
$$T = \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \dots$$
According to the Tortoise, Achilles will never catch up, and thus the time T at which Achilles overtakes is (informally) $T = +\infty$. However, a simple algebraic argument proved that $T = 1$.

The flaw in the Tortoise's argument is therefore this:

> The Tortoise assumes that a "sum" with infinitely many strictly positive terms must necessarily add up to infinity.

Because we know that the Tortoise is wrong, we now have valuable information: *It must be possible to add up infinitely many strictly positive terms, and yet end up with a finite answer.* However, addition is a binary operation: You can add up only two numbers at a time. Thus, for us, an expression such as
$$\frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \dots$$
has, *as yet*, no meaning. We need to *provide* it with a meaning. Moreover, for that meaning to be consistent with our algebraic argument, our definition must ensure that
$$\frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \dots = 1$$

[2] In Zeno's version, Achilles concedes the race, dazzled by the Tortoise's logic.

In the previous chapter we proved that every real number has a decimal representation. It should be clear that, in decimal notation,

$$\frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \cdots = 0.9999\ldots$$

But (of course) $0.9999\cdots = 9 \times (0.1111\ldots) = 9 \times \frac{1}{9} = 1$, so there is no contradiction here.

Exercise 2.5.1 Here's another one of Zeno's paradoxes: *Motion is impossible.*
For suppose that I want to walk a distance of $x > 0$ meters. Before I get to the x meter mark, I would first have to reach the $\frac{1}{2}x$ meter mark. And before I can reach the $\frac{1}{2}x$ meter mark, I would first have to reach the $\frac{1}{4}x$ meter mark...
Thus there are infinitely many distances I have to travel before I reach the point x, and doing infinitely many things will take me an infinite amount of time.
Examine this argument, and point out the flaws.

□

2.5.2 Convergence of Series: Definition and Examples

Note that, as written, the expression

$$x_1 + x_2 + x_3 + \cdots + x_n + \ldots \qquad (*)$$

seems to require an infinite number of additions. We have no idea what that might mean, because addition is a *binary* operation — we can add only two numbers at a time. We therefore need to find a clear and unambiguous meaning for $(*)$.

A careful examination of the resolution of Zeno's paradoxes leads us to believe that it may be possible provide such a meaning.

We shall do this as follows: For each n, define

$$s_n = x_1 + x_2 + \cdots + x_n = \sum_{k=1}^{n} x_k$$

Each s_n is called a *partial sum* of the series, and is a well–defined number, as it involves only finitely many additions. We thus obtain a sequence $\langle s_n \rangle_n$ of partial sums. We shall *define* the series $(*)$ to be the sequence $\langle s_n \rangle_n$.
Thus a series is a sequence, not a number!

In particular,

$$\text{The series} \quad \sum_{k=1}^{\infty} x_k \quad \textbf{\textit{is}} \text{ the sequence} \quad \Big\langle \sum_{k=1}^{n} x_k \Big\rangle_n$$

If the series of partial sums converges, i.e. if $s_n \to s$, then we shall say that the series converges to s, and write

$$\sum_{k=1}^{\infty} x_k = s \quad \text{instead of} \quad \lim_{n\to\infty} \sum_{k=1}^{n} x_k = s$$

Thus when we say $\sum_{k=1}^{\infty} x_k = s$, we mean that the sequence of partial sums converges, and that its limit is s. A series which does not converge is said to diverge.

> *A series converges to s if and only if you can get as close to s as you like by adding up sufficiently many terms of the series.*

Some remarks on notation: We will write $\sum_{k=m}^{\infty} x_k$ for the series $x_m + x_{m+1} + x_{m+2} + \ldots$. We may also write $\sum_n x_n$ instead of $\sum_{n=1}^{\infty} x_n$, if there is no danger of confusion.

Furthermore, we write $\sum_n x_n = \infty$ if $s_n \to \infty$. But note that in that case the series is *divergent*.

Example 2.5.2 Let $x \in \mathbb{R}$ with $|x| < 1$. Then the series $\sum_{k=0}^{\infty} x^k$ is the sequence $\langle s_n \rangle_n$, where

$$s_n = 1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x}$$

Now $x^{n+1} \to 0$ as $n \to \infty$, because $|x| < 1$ (see Example 2.3.7). It follows that $s_n \to \frac{1}{1-x}$. We therefore write

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x}$$

This does *not* mean that if you add up all the (infinitely many) x^k, you end up with $\frac{1}{1-x}$ — you can't add up infinitely many terms. Instead, it means that the sequence $\langle s_n \rangle_n$ of partial sums converges to $\frac{1}{1-x}$.

□

Example 2.5.3 The series

$$1 - 1 + 1 - 1 + 1 - \ldots$$

diverges. For

$$s_n = \begin{cases} 1 \text{ if } n \text{ is odd} \\ 0 \text{ if } n \text{ is even} \end{cases}$$

i.e. $\langle s_n \rangle_n$ is the sequence $1, 0, 1, 0, \ldots$, and that diverges.

□

Combining Examples 2.5.2 and 2.5.3 leads to the following result:

Theorem 2.5.4 *The series $\sum_{n=0}^{\infty} x^n$ converges if and only if $|x| < 1$.*
When $|x| < 1$ we have

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x}$$

Example 2.5.5 (Harmonic Series)
We prove the following important fact:

$$\text{The } \textit{harmonic series} \quad \sum_{n=1}^{\infty} \frac{1}{n} \quad \text{diverges.}$$

We show that the sequence $\langle s_n \rangle_n$ of partial sums is unbounded. As every convergent sequence is bounded (cf. proposition 2.2.22), this implies that $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. We shall accomplish this by proving that $s_{2^n} \geq 1 + \frac{n}{2}$ for all n.

We consider partial sums of the form s_{2^n} (i.e. $s_1, s_2, s_4, s_8, \dots$) because we can group their terms in an ingenious way to obtain a lower bound:

$$\begin{aligned} s_{2^n} = 1 + \frac{1}{2} + \Big(\frac{1}{3} + \frac{1}{4}\Big) + \Big(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\Big) + \dots \\ + \Big(\frac{1}{2^{n-1}+1} + \frac{1}{2^{n-1}+2} + \dots + \frac{1}{2^n}\Big) \end{aligned}$$

Thus after the first two terms (1 and $\frac{1}{2}$), we group the remaining terms in brackets: The next two ($\frac{1}{3}$ and $\frac{1}{4}$), then the next four, then the next eight, etc., until we get a final bracket with 2^{n-1} terms. There are $n-1$ brackets in total.

Now clearly

$$\begin{aligned} \frac{1}{3} + \frac{1}{4} &\geq \frac{1}{4} + \frac{1}{4} \\ \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} &\geq \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \\ \frac{1}{2^{n-1}+1} + \frac{1}{2^{n-1}+2} + \dots + \frac{1}{2^n} &\geq \frac{1}{2^n} + \frac{1}{2^n} + \dots + \frac{1}{2^n} \end{aligned}$$

Now each expression on the righthand side adds up to exactly $\frac{1}{2}$, and there are $n-1$ such expressions. It follows that

$$s_{2^n} \geq 1 + \frac{1}{2} + \frac{1}{2}(n-1) = 1 + \frac{n}{2}$$

Thus the sequence $\langle s_n \rangle_n$ is unbounded, which means that the series diverges.

□

Example 2.5.6 We show that if $p > 1$, then the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges. The proof is very similar to the argument presented in the previous example. We split up partial sums of the form s_{2^n-1} into groups with power-of-two-many terms:

$$\begin{aligned} s_{2^n-1} = 1 + \Big(\frac{1}{2^p} + \frac{1}{3^p}\Big) + \Big(\frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p}\Big) + \dots \\ + \Big(\frac{1}{2^{(n-1)p}} + \frac{1}{(2^{n-1}+1)^p} + \dots + \frac{1}{(2^n-1)^p}\Big) \end{aligned}$$

Now

$$\begin{aligned} \frac{1}{2^p} + \frac{1}{3^p} &\leq \frac{1}{2^p} + \frac{1}{2^p} = \frac{1}{2^{p-1}} \\ \frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p} &\leq \frac{1}{4^p} + \frac{1}{4^p} + \frac{1}{4^p} + \frac{1}{4^p} = \Big(\frac{1}{2^{p-1}}\Big)^2 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \\ \frac{1}{2^{(n-1)p}} + \frac{1}{(2^{n-1}+1)^p} + \dots + \frac{1}{(2^n-1)^p} &\leq \frac{1}{2^{(n-1)p}} + \frac{1}{2^{(n-1)p}} + \dots + \frac{1}{2^{(n-1)p}} = \Big(\frac{1}{2^{p-1}}\Big)^{n-1} \end{aligned}$$

It follows that we have majorized the sequence s_{2^n-1} by a finite geometric series:

$$s_{2^n-1} \leq 1 + \frac{1}{2^{p-1}} + \left(\frac{1}{2^{p-1}}\right)^2 + \cdots + \left(\frac{1}{2^{p-1}}\right)^{n-1} \leq \frac{1 - \left(\frac{1}{2^{p-1}}\right)^n}{1 - \frac{1}{2^{p-1}}}$$

Since $p > 1$, we have $\left(\frac{1}{2^{p-1}}\right)^n \to 0$ (as $n \to \infty$), and so $\frac{1-\left(\frac{1}{2^{p-1}}\right)^n}{1-\frac{1}{2^{p-1}}} \leq \frac{1}{1-\frac{1}{2^{p-1}}}$. It is now easy to see that the sequence $\langle s_n \rangle_n$ is bounded: Given $k \in \mathbb{N}$, choose an n such that $k \leq 2^n - 1$. Then $s_k \leq s_{2^n-1}$ (because all terms are positive, and s_{2^n-1} includes all the terms of s_k, and possibly more). Moreover, $s_{2^n-1} \leq \frac{1}{1-\frac{1}{2^{p-1}}}$. We have therefore shown that

$$s_k \leq \frac{1}{1 - \frac{1}{2^{p-1}}} \quad \text{for all } k \in \mathbb{N}$$

and thus that the sequence $\langle s_n \rangle_n$ is bounded above. Since it is also an increasing sequence, it must converge, by Theorem 2.3.11.

Thus the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges when $p > 1$.

□

Combining Examples 2.5.5 and 2.5.6 leads to the following result:

Theorem 2.5.7 *The series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if $p > 1$, and diverges if $p \leq 1$.*

Proof: Examples 2.5.5 and 2.5.6 supplied proofs for the cases $p = 1$ and $p > 1$ respectively. If $p < 1$, then $\frac{1}{n^p} \geq \frac{1}{n}$. Since the partial sums of $\sum_{n=1}^{\infty} \frac{1}{n}$ are unbounded, it follows easily that the partial sums of $\sum_{n=1}^{\infty} \frac{1}{n^p}$ are unbounded as well. Since a convergent sequence *must* be bounded (cf. Propn. 2.2.22), and since a series *is* sequence of partial sums, we see that $\sum_{n=1}^{\infty} \frac{1}{n^p}$ diverges.

⊣

Remarks 2.5.8 In Section 2.4, we proved that every number has a decimal representation. We first defined what we *meant* by an expression of the form $0.a_1a_2a_3\ldots$, and decided that we should define it to be $\lim_n s_n$, where $s_n = \frac{a_1}{10} + \frac{a_2}{100} + \cdots + \frac{a_n}{10^n}$. Thus we defined $0.a_1a_2a_3\ldots$ to be $\sum_n \frac{a_n}{10^n}$, i.e. the decimal representation of a number is a series.

□

Exercise 2.5.9 Suppose that $\langle x_n \rangle_n$ and $\langle y_n \rangle_n$ are sequence in $\mathbb{R}$, and that $\alpha \in \mathbb{R}$.

(a) If $\sum_n x_n = x$ and $\sum_n y_n = y$, then $\sum_n (x_n + y_n) = x + y$.

(b) If $\sum_n x_n = x$, then $\sum_n \alpha x_n = \alpha x$.

[Hint: (a) Remember that series are sequences of partial sums. So let $s_n = \sum_{k=1}^n x_k, t_n = \sum_{k=1}^n y_k, u_n = \sum_{k=1}^n (x_k + y_k)$. You must show that $u_n \to (x + y)$, given that $s_n \to x$ and $t_n \to y$. Use Theorem 2.3.1.]

□

Exercise 2.5.10 (The number e)
Show that the series

$$\sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots$$

converges. The number e is defined to be the limit of this series, i.e.

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

[Hint: $n! \geq 2^{n-1}$ if $n \geq 2$. Use this fact to show that the partial sums of $\sum_{n=0}^{\infty} \frac{1}{n!}$ form an increasing bounded sequence, and invoke Theorem 2.3.11.]

□

2.6 Convergence of Subsequences

2.6.1 Subsequences

Roughly speaking, if you write down all the terms of a sequence $\langle x_n \rangle_n$, and then delete some of these terms, what remains is a subsequence. However, you're not allowed to delete so many terms that only finitely remain, nor are you allowed to rearrange the order in which they occur.
This is best understood by looking at some examples: The sequence $2, 3, 5, 7, 11, \dots$ of primes is a subsequence of the sequence $1, 2, 3, 4, \dots$ of natural numbers:

$$\not{1}, 2, 3, \not{4}, 5, \not{6}, 7, \not{8}, \not{9}, \not{1}0, 11, \dots$$

In the subsequence, the order of elements remains the same as what it was in the original: 2 comes before 3 comes before 5... etc. in both sequences.
The sequence $3, 2, 6, 5, 9, 8, \dots$ is *not* a subsequence of $1, 2, 3, 4, 5, \dots$. Not only have we deleted all numbers of the form $3n - 2$, we have also *rearranged* them so that $3n$ is before $3n-1$. In the sequence of natural numbers, 2 is before 3, but in this new sequence, 3 is before 2. Such rearrangements are not allowed when you construct a subsequence.

The following definition should now make sense:

Definition 2.6.1 Let $\langle x_n \rangle_n$ be a sequence in $\mathbb{R}$, and suppose that $\langle n_k \rangle_k$ is a strictly increasing sequence in $\mathbb{N}$ (i.e. $n_1 < n_2 < n_3 < \dots$). Then the sequence

$$\langle x_{n_k} \rangle_k = x_{n_1}, x_{n_2}, x_{n_3}, \dots$$

is called a *subsequence* of $\langle x_n \rangle_n$.

For example

$$\langle x_{2n} \rangle_n = x_2, x_4, x_5, \dots$$
$$\langle x_{3n-1} \rangle_n = x_2, x_5, x_8, \dots$$
$$\langle x_{5^n} \rangle_n = x_5, x_{25}, x_{125}, \dots$$

are subsequences of $\langle x_n \rangle_n$.

Remarks 2.6.2 1. One easy but useful fact to note is the following: If $n_1 < n_2 < n_3 < \dots$ is a strictly increasing sequence of natural numbers, then $n_k \geq k$ (for each $k \in \mathbb{N}$).
If you can't see this immediately, try proving it by induction. Clearly $n_1 \geq 1$. Now suppose that $n_k \geq k$. Then $n_{k+1} > n_k \geq k$, and thus $n_{k+1} \geq k+1$.

2. Note that the n in $\langle x_n \rangle_n$ is a "dummy" variable — not really a variable at all. This means that it doesn't matter if we replace the n by some other symbol k: $\langle x_k \rangle_k$ is *exactly the same* as $\langle x_n \rangle_n$.
For example $\langle \frac{1}{k} \rangle_k = 1, \frac{1}{2}, \frac{1}{3}, \dots = \langle \frac{1}{n} \rangle_n$.
In particular, $\lim_k x_k$ is exactly the same as $\lim_n x_n$, $\sup_k x_k$ the same as $\sup_n x_n$, etc.

In the expression $\langle x_n \rangle_n$, the variable n is a *bound* variable, constrained to take on *all* possible values in the set $\mathbb{N}$. We have a similar situation when we deal with definite integrals: The expression $\int_0^1 x \, dx$ is a number, namely $\frac{1}{2}$, and not a variable, even though it seems to have a variable x occurring in it. However, *that* x is a bound variable, constrained to take on all possible values between 0 and 1. It doesn't matter if we replace the$x$ by some other symbol u: $\int_0^1 x \, dx$ is *exactly the same* as $\int_0^1 u \, du$

□

One important type of subsequence is a *tail sequence*. A tail sequence of $\langle x_n \rangle_n$ is a subsequence which consists of all terms of x_n from some N onwards, e.g. $5, 6, 7, \dots$ is a tail sequence of $1, 2, 3, \dots$. Similarly $\frac{1}{100}, \frac{1}{101}, \frac{1}{102}, \dots$ is a tail sequence of $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$. Thus $\langle y_n \rangle_n$ is a tail sequence of $\langle x_n \rangle_n$ iff there is an integer $N \geq 0$ such that $y_n = x_{N+n}$.

Example 2.6.3 If

$$x_n = \begin{cases} \dfrac{1}{n} & \text{if } n \text{ is odd} \\ 1 + \dfrac{1}{n^2} & \text{if } n \text{ is even} \end{cases}$$

then $\langle x_n \rangle_n$ is divergent. However, the sequences $\langle y_n \rangle_n, \langle z_n \rangle_n$ defined by

$$y_n = x_{2n-1}, \qquad z_n = x_{2n}$$

are convergent subsequences of $\langle x_n \rangle_n$, with $y_n \to 0$, and $z_n \to 1$.
If you think long enough, it should be clear that a subsequence $\langle x_{n_k} \rangle_k$ of $\langle x_n \rangle_n$ converges if and only if: EITHER the sequence $\langle n_k \rangle_k$ is odd eventually (in which case $x_{n_k} \to 0$), OR $\langle n_k \rangle_k$ is even eventually (in which case $x_{n_k} \to 1$).
Similarly, if $\langle n_k \rangle_k$ is BOTH odd infinitely often and even infinitely often, then $\langle x_{n_k} \rangle_k$ diverges.

□

Proposition 2.6.4 *(a) If $x_n \to x$, and if $\langle y_n \rangle_n$ is a subsequence of $\langle x_n \rangle_n$, then $y_n \to x$ as well.*

(b) If $\langle y_n \rangle_n$ is a tail sequence of $\langle x_n \rangle_n$, and if $y_n \to x$, then also $x_n \to x$.

Proof: (a) Suppose that $y_k = x_{n_k}$, where $n_1 < n_2 < n_3 < \dots$. We must show that $y_k \to x$, i.e. that for every $\varepsilon > 0$, there is a $K \in \mathbb{N}$ such that $|y_k - x| < \varepsilon$ whenever $k > K$.

So let $\varepsilon > 0$ be given. First choose $M \in \mathbb{N}$ such that $|x_m - x| < \varepsilon$ whenever $m > M$. (Why can we choose such an M?)

Now choose $K \in \mathbb{N}$ such that $n_k > M$ whenever $k > K$. (For example, by Remarks 2.6.2, we have $n_k \geq k$ for all $k \in \mathbb{N}$. In particular, $n_M \geq M$, so we can choose $K = M$.) Then if $k > K$, we have

$$|y_k - x| = |x_{n_k} - x| < \varepsilon$$

(because $n_k > M$ implies $|x_{n_k} - x| < \varepsilon$). Thus $y_k \to x$, as required.

(b) Suppose that $\langle y_n \rangle_n$ is a convergent tail subsequence of $\langle x_n \rangle_n$, and that $y_n \to x$. We must show that also $x_n \to x$. So let $\varepsilon > 0$. Choose N such that $n > N$ implies $|y_n - x| < \varepsilon$. Next, note that that by definition of "tail sequence", there is a non–negative integer M is such that $y_n = x_{n+M}$. It follows that if $n > N + M$, then $n - M > N$, so that $|y_{n-M} - x| < \varepsilon$. But $y_{n-M} = x_n$, and thus

$$|x_n - x| < \varepsilon \quad \text{whenever} \quad n > N + M$$

Since we can do this for any $\varepsilon > 0$, we have shown that $x_n \to x$.

⊣

Example 2.6.5 Consider again the sequence

$$x_n = \begin{cases} \dfrac{1}{n} & \text{if } n \text{ is odd} \\ 1 + \dfrac{1}{n^2} & \text{if } n \text{ is even} \end{cases}$$

The sequences

$$y_n = \frac{1}{2n+1} \qquad z_n = 1 + \frac{1}{4n^2}$$

are subsequences of $\langle x_n \rangle_n$. Since $y_n \to 0$ and $z_n \to 1$, we can conclude that $\langle x_n \rangle_n$ is divergent. For if $\langle x_n \rangle_n$ converges (to x, say), then all its subsequences would also converge to the same limit x. But here we have two subsequences which converge to different limits.

□

Next, we show that every sequence of real numbers has a monotone subsequence. For the purpose of the proof, we briefly introduce some non–standard terminology. Let $\langle x_n \rangle_n$ be a sequence of real numbers. Imagine that you are walking along a landscape, and that x_n is your height above sea level at time n. Call x_n a *vista* if you can see the whole landscape ahead of you, i.e. if $x_n \geq x_m$ for all $m \geq n$. Thus if $\langle x_n \rangle_n$ is decreasing, then each x_n is a vista, whereas if $\langle x_n \rangle_n$ is increasing, there are no vistas at all. If $x_n := 1 + (-1)^n \frac{1}{n}$, then every even point x_{2n} is a vista.

Theorem 2.6.6 *Every sequence of real numbers has a monotone subsequence.*

Proof: We consider two cases: Either (1) $\langle x_n \rangle_n$ has infinitely many vistas, or (2) it has only finitely many. In case (1), let $x_{n_1}, x_{n_2}, x_{n_3}, \ldots$ be the subsequence of vistas, in order of increasing subscript. Note that

$$x_{n_1} \geq x_{n_2} \geq x_{n_3} \geq \cdots \geq x_{n_k} \geq \ldots$$

is a decreasing subsequenc of $\langle x_n \rangle_n$.

In case (2), $\langle x_n \rangle_n$ has only finitely many vistas, so there is an $N \in \mathbb{N}$ such that there are no vistas beyond point x_N, i.e. if $n \geq N$, then x_n is not a vista. Now construct a subsequence as follows. Let $n_1 = N$. Since x_{n_1} is not a vista, there is $n_2 > n_1$ such that $x_{n_2} > x_{n_1}$. Since x_{n_2} is not a vista, there is $n_3 > n_2$ such that $x_{n_3} > x_{n_2}$. Continuing in this way, we obtain an increasing sequence

$$x_{n_1} < x_{n_2} < x_{n_3} \ldots x_{n_k} < \ldots$$

⊣

2.6.2 Bolzano–Weierstrass Theorem

This theorem is so important that it deserves a subsection all to itself:

Theorem 2.6.7 (Bolzano–Weierstrass) *Every bounded sequence of real numbers has a convergent subsequence.*

Proof: By Theorem 2.6.6, any sequence $\langle x_n \rangle$ has a monotone subsequence. If $\langle x_n \rangle$ is bounded, then so is the subsequence. But a bounded monotone sequence converges, by Theorem 2.3.11.

⊣

2.7 Cauchy Sequences and Completeness

We have already seen that any bounded increasing sequence converges (Theorem 2.3.11) — a fact that followed from the Completeness Axiom. This fact allowed us, in Example 2.3.13, to conclude that the sequence $\langle (1+\frac{1}{n})^n \rangle_n$ converges, though we could not see where it converges to. The Completeness Axiom guarantees the existence of a limit, even if we do not know what that limit is.

Like a bounded increasing sequence, a *Cauchy sequence* is a sequence that "ought to" converge. And, as we shall see, a Cauchy sequence *does* converge: The existence of a limit is guaranteed by the Completeness Axiom, even if we do not know what that limit actually is.

Intuitively, a sequence $\langle x_n \rangle_n$ in $\mathbb{R}$ is a Cauchy sequence if its terms lie *eventually arbitrarily close* to each other. This means that from some point onwards, any two terms are "close". If all terms lie closer and closer together, there should be some point that they are all clustering around, and that point should be the limit of the sequence $\langle x_n \rangle_n$.

All this "ought" and "should" needs to be made precise.

Definition 2.7.1 A sequence $\langle x_n \rangle_n$ in $\mathbb{R}$ is called a *Cauchy sequence* if and only if for every $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$|x_n - x_m| < \varepsilon \quad \text{whenever} \quad n, m \geq N$$

i.e. if and only if

$$\forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n, m \geq N \; [|x_n - x_m| < \varepsilon]$$

Remarks 2.7.2 (a) Note that *all* terms from some point onwards need to be within ε of each other, not just successive terms. Thus, for example, if $N = 100$, then not just do we have $|x_{100} - x_{101}| < \varepsilon$, but also $|x_{301} - x_{15\,673\,428}| < \varepsilon$.

(b) A neat way to characterize Cauchy sequences is as follows:

$$\langle x_n\rangle_n \text{ is a Cauchy sequence} \qquad \Longleftrightarrow \qquad \lim_{N\to\infty} \sup_{n\geq N} |x_n - x_N| = 0$$

Here ($\Rightarrow$) is obvious. ($\Leftarrow$) follows by the triangle inequality: Given $\varepsilon > 0$, choose N such that $\sup_{k\geq N} |x_k - x_N| < \varepsilon/2$. Then for $n, m \geq N$ we have

$$|x_n - x_m| \leq |x_n - x_N| + |x_N - x_m| \leq 2 \sup_{k\geq N} |x_k - x_N| < \varepsilon$$

□

Example 2.7.3 The sequence $\langle 1 + (-1)^n 2^{-n}\rangle_n$ is Cauchy. Indeed, given $\varepsilon > 0$, we may choose $N \in \mathbb{N}$ such that $2^{-N} < \frac{\varepsilon}{2}$. If $n, m > N$, then (by the triangle inequality)

$$\left|(1 + (-1)^n 2^{-n}) - (1 + (-1)^m 2^{-m})\right| \leq 2^{-n} + 2^{-m} \leq 2^{-N} + 2^{-N} < \varepsilon$$

□

Lemma 2.7.4 *Every convergent sequence is a Cauchy sequence*

Proof: Suppose that $x_n \to x$, and that we are given $\varepsilon > 0$. We must find N such that $|x_n - x_m| < \varepsilon$ whenever $n, m > N$.

Now because $x_n \to x$ there is $N \in \mathbb{N}$ such that $|x_n - x| < \frac{\varepsilon}{2}$ whenever $n \geq N$. In particular, if $n, m \geq N$, then

$$|x_n - x_m| \leq |x_n - x| + |x - x_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

Hence $\langle x_n\rangle_n$ is a Cauchy sequence.

⊣

So any convergent sequence is a Cauchy sequence. And this is not surprising: If the terms of a sequence $\langle x_n\rangle_n$ are eventually close to some point x (the limit), then those terms must also eventually be close to each other.

More importantly, the converse is true: Any Cauchy sequence in $\mathbb{R}$ is convergent. To prove this, we will need a number of lemmas. We shall prove:

- Every Cauchy sequence is bounded.
- Every bounded sequence has a convergent subsequence.
- If a Cauchy sequence $\langle x_n\rangle_n$ has a convergent subsequence, then $\langle x_n\rangle_n$ is itself convergent.

Actually, the second point has already been proved. It is the Bolzano–Weierstrass theorem (Theorem 2.6.7). Thus we need only prove the first and the last point.

Lemma 2.7.5 *If $\langle x_n\rangle_n$ is a Cauchy sequence in $\mathbb{R}$, then $\langle x_n\rangle_n$ is bounded.*

Proof: Choose $N \in \mathbb{N}$ such that $|x_n - x_m| < 1$ whenever $n, m \geq N$. (This is possible, because $\langle x_n \rangle_n$ is Cauchy — we have taken $\varepsilon = 1$.) Now define

$$K = \max\{|x_1|, |x_2|, \ldots, |x_N| + 1\}$$

We show that K is a bound for $\langle x_n \rangle_n$, i.e. that $|x_n| \leq K$ for *all* $n \in \mathbb{N}$.

Consider separately the two case (i) $n \leq N$, and (ii) $n > N$. In case (i), we obviously have $|x_n| \leq K$, by definition of K. Suppose therefore, that $n > N$. In that case, both n and N are $\geq N$, and thus

$$|x_n| \leq |x_n - x_N| + |x_N| \leq 1 + |x_N| \leq K$$

which finishes case (ii).

⊣

Lemma 2.7.6 *If $\langle x_n \rangle_n$ is a Cauchy sequence, and if $\langle x_n \rangle_n$ has a convergent subsequence, then $\langle x_n \rangle_n$ itself converges.*

Proof: Suppose that $\langle x_{n_k} \rangle_k$ is a subsequence of the Cauchy sequence $\langle x_n \rangle_n$, and that $x_{n_k} \to x$ (as $k \to \infty$). We show that $x_n \to x$ (as $n \to \infty$).

So let $\varepsilon > 0$. We must show that there is $N \in \mathbb{N}$ such that $|x_n - x| < \varepsilon$ whenever $n \geq N$. Now because $\langle x_n \rangle_n$ is a Cauchy sequence, we can find an N_1 such that

$$n, m \geq N_1 \quad \text{implies} \quad |x_n - x_m| < \tfrac{\varepsilon}{2}$$

Because $x_{n_k} \to x$, we can find a K such that

$$k \geq K \quad \text{implies} \quad |x_{n_k} - x| < \tfrac{\varepsilon}{2}$$

Now define $N = \max\{N_1, n_K\}$, and let $n \geq N$. Choose k such that $n_k \geq N$. Then (i) $n, n_k \geq N_1$, and (ii) $k \geq K$ (because $n_k \geq N \geq n_K$). It follows that

$$|x_n - x| \leq |x_n - x_{n_k}| + |x_{n_k} - x| < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2}$$

whenever $n > N$.

⊣

Theorem 2.7.7 *Let $\langle x_n \rangle_n$ be a sequence in $\mathbb{R}$. Then $\langle x_n \rangle_n$ converges if and only if it is a Cauchy sequence.*

Proof: ($\Rightarrow$) is Lemma 2.7.4.
($\Leftarrow$): If $\langle x_n \rangle_n$ is a Cauchy sequence, then it is bounded (by Lemma 2.7.5). Hence it has a convergent subsequence (by Theorem 2.6.7). It follows that $\langle x_n \rangle_n$ converges (by Lemma 2.7.6).

⊣

Remarks 2.7.8 The fact that Cauchy sequence converge in $\mathbb{R}$ is depends very much on the Completeness Axiom. If you look back over the proof of Theorem 2.7.7, you will not see the Completeness Axiom mentioned explicitly. But we *do* use the Bolzano–Weierstrass Theorem. The latter's proof depended on the fact that bounded monotone sequences converge, and that fact, in turn, requires the Completeness Axiom.

□

Exercises 2.7.9 1. (a) Prove that if $\langle x_n \rangle$ converges, then $\lim_n(x_{n+1} - x_n) = 0$.

(b) Does the converse hold? i.e., does $\lim_n(x_n - x_{n+1}) = 0$ imply that $\langle x_n \rangle$ converges?

2. (a) Suppose that a sequence $\langle x_n \rangle$ has $|x_{n+1} - x_n| \leq 2^{-n}$ for all $n \in \mathbb{N}$. Show that $\langle x_n \rangle$ converges.

(b) Does the same hold if we only know that $|x_n - x_{n+1}| \leq \frac{1}{n}$ for all $n \in \mathbb{N}$?

□

Exercise 2.7.10 Suppose that $a < b$ and that $0 < \lambda < 1$. Let $\langle x_n \rangle$ be a sequence of real numbers defined inductively as follows:

$$x_1 = a, \quad x_2 = b, \qquad x_{n+1} = \lambda x_{n-1} + (1 - \lambda)x_n \quad \text{for } n \geq 2$$

(a) Show that $|x_{n+1} - x_n| = \lambda |x_n - x_{n-1}|$.

(b) Conclude that $|x_{n+1} - x_n| = \lambda^{n-1}(b - a)$

(c) Prove that if $n > m$, then $|x_n - x_m| \leq (b - a) \cdot \lambda^{m-1} \sum_{k=0}^{n-m-1} \lambda^k$.
[Hint: Triangle inequality.]

(d) Deduce that $|x_n - x_m| \leq \frac{\lambda^{m-1}(b-a)}{1-\lambda}$ when $n \geq m$.

(e) Now prove that $\langle x_n \rangle$ converges by showing that it is a Cauchy sequence.

□

2.8 Further Results on Convergence of Series

2.8.1 Cauchy Criteria

The following result is often useful:

Theorem 2.8.1 (Cauchy Criterion)
The series $\sum_{n=1}^{\infty} x_n$ converges if and only if

$$\textit{For every } \varepsilon > 0 \textit{ there is } N \in \mathbb{N} \textit{ such that } m > n \geq N \textit{ implies } \Big| \sum_{k=n+1}^{m} x_k \Big| < \varepsilon$$

Proof: Recall that a sequence converges if and only if it is a Cauchy sequence (Theorem 2.7.7). Now the series $\sum_{n=1}^{\infty} x_n$ *is* the sequence $\langle s_n \rangle_n$ of partial sums, and will therefore converge if and only if $\langle s_n \rangle_n$ is a Cauchy sequence. Now note that if $m \geq n$, then

$$|s_m - s_n| = \Big| \sum_{k=n+1}^{m} x_k \Big|$$

and we can make $|s_m - s_n|$ as small as we like by taking m, n to be sufficiently large.

⊣

Taking $m = k$, $n = k - 1$ in the preceding theorem, we see that $|s_m - s_n| = |x_k| < \varepsilon$ whenever $k > N$. It follows immediately that:

Corollary 2.8.2 *If $\sum_{n=1}^{\infty} x_n$ converges, then $x_n \to 0$.*

Alternatively, if $\sum_{n=1}^{\infty} x_n = s$, then the sequence of partial sums has $s_n \to s$, and thus $\lim_n x_n = \lim_n (s_n - s_{n-1}) = \lim_n s_n - \lim_n s_{n-1} = s - s = 0$.

Note that the converse of Corollary 2.8.2 is not true — see Example 2.5.5 on the divergence of the harmonic series. However, we can say something about convergence if the terms of the series alternate in sign. An example of an alternating series is

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

Theorem 2.8.3 (Alternating Series) *Suppose that $\langle x_n \rangle_n$ is a decreasing non–negative sequence in $\mathbb{R}$ such that $x_n \to 0$. Then the alternating series*

$$\sum_{n=1}^{\infty} (-1)^{n+1} x_n = x_1 - x_2 + x_3 - x_4 + \dots$$

converges.

Proof: Note that, since $\langle x_n \rangle_n$ is decreasing, we have $x_k - x_{k+1} \geq 0$, for all $k \in \mathbb{N}$. It follows that, if $m > n$, then

$$0 \leq x_{n+1} - x_{n+2} + x_{n+3} - \dots \pm x_m \leq x_{n+1}$$

We now apply the Cauchy criterion. As always, let $s_n = \sum_{k=1}^{n} (-1)^{k+1} x_k$, and let $\varepsilon > 0$. Since $x_n \to 0$, there is $N \in \mathbb{N}$ such that $x_n < \varepsilon$ whenever $n > N$. It follows that if $m \geq n > N$, then

$$|s_m - s_n| = |x_{n+1} - x_{n+2} + x_{n+3} - \dots \pm x_m| \leq x_{n+1} < \varepsilon$$

Thus $\langle s_n \rangle_n$ is a Cauchy sequence, and thus convergent.

⊣

Example 2.8.4 It follows from Theorem 2.8.3 that the alternating harmonic series

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

converges.

□

The method of Examples 2.5.5 and 2.5.6 can be generalized. We state here the following result:

Theorem 2.8.5 (Cauchy Condensation Test)
Suppose that $\langle x_k \rangle_k$ *is a* decreasing non–negative *sequence. Then the series* $\sum_{k=1}^{\infty} x_k$ *converges if and only if the series*

$$\sum_{k=0}^{\infty} 2^k x_{2^k} = x_1 + 2x_2 + 4x_4 + 8x_8 + \dots$$

converges.

For example, $\sum_{n=1}^{\infty} \frac{1}{n}$ converges if and only if $\sum_{n=0}^{\infty} 2^n \frac{1}{2^n}$ converges. But the latter series obviously diverges, and thus the harmonic series diverges as well.

The proof of Theorem 2.8.5 is left as an exercise:

Exercise 2.8.6 We prove Theorem 2.8.5.
Let $\langle s_n \rangle_n$ and $\langle t_n \rangle_n$ be the sequences of partial sums of the above series, i.e.

$$s_n = \sum_{k=1}^{n} x_k \qquad\qquad t_n = \sum_{k=0}^{n} 2^k x_{2^k}$$

It is clear that $\langle s_n \rangle_n$ and $\langle t_n \rangle_n$ are increasing sequences. We must show that $\langle s_n \rangle_n$ converges if and only if $\langle t_n \rangle_n$ converges.

(a) Explain why it suffices to show that $\langle s_n \rangle_n$ is bounded if and only if $\langle t_n \rangle_n$ is bounded.

(b) Suppose that $n < 2^k$. Explain why

$$s_n \leq x_1 + (x_2 + x_3) + (x_4 + x_5 + x_6 + x_7) + \dots + (x_{2^k} + \dots + x_{2^{k+1}-1}) \leq t_k$$

(c) Deduce that if $\langle s_n \rangle_n$ is unbounded, then so is $\langle t_n \rangle_n$.

(d) Next, let $n > 2^k$. Explain why

$$s_n \geq x_1 + x_2 + (x_3 + x_4) + (x_5 + x_6 + x_7 + x_8) + \dots + (x_{2^{k-1}+1} + \dots + x_{2^k}) \geq \frac{1}{2} t_k$$

(e) Deduce that if $\langle t_n \rangle_n$ is unbounded, then so is $\langle s_n \rangle_n$.

(f) Now explain why $\sum_{k=1}^{\infty} x_k$ converges if and only if $\sum_{k=0}^{\infty} 2^k x_{2^k}$ converges.

□

Here's some more practice in the use of Theorem 2.8.5:

Exercise 2.8.7 Use Theorem 2.8.5 to prove that $\sum_{n=1}^{\infty} \frac{1}{n(\ln n)^p}$ converges if $p > 1$, and diverges if $p \leq 1$.

□

If $\sum_n x_n$ is a series, then by a *tail series* we mean a series of the form $\sum_{n=N+1}^{\infty} x_n$. Note the following rather obvious, but useful, facts:

Proposition 2.8.8 *(a) Let $\sum_n x_n$ be a series, and let $N \in \mathbb{N}$. Then $\sum_n x_n$ converges if and only if the tail series $\sum_{n=N+1}^{\infty} x_n$ converges. In that case*

$$\sum_{n=1}^{\infty} x_n = \sum_{n=1}^{N} x_n + \sum_{n=N+1}^{\infty} x_n$$

(b) The series $\sum_n x_n$ converges if and only if the sequence of tail series $\langle \sum_{k=n+1}^{\infty} x_k \rangle_n$ converges to zero.

Proof: (a) Let s_n be the n^{th} partial sum of $\sum_n x_n$, and t_n the n^{th} partial sum of $\sum_{n=N+1}^{\infty} x_n = \sum_{n=1}^{\infty} x_{N+n}$, i.e.

$$s_n = x_1 + x_2 + \cdots + x_n \qquad t_n = x_{N+1} + x_{N+2} + \cdots + x_{N+n}$$

Note that $s_{N+n} = s_N + t_n$. Now the sequence $\langle s_n \rangle_n$ converges if and only if its tail $\langle s_{N+n} \rangle_n$ converges, in which case they converge to the same limit — cf. Proposition 2.6.4. Since $s_{N+n} = s_N + t_n$, we see that $\langle s_n \rangle_n$ converges if and only if $\langle t_n \rangle_n$ converges, and thus that $\sum_n x_n$ converges if and only if $\sum_{n=N+1}^{\infty} x_n$ converges. Moreover, if $\sum_n x_n = x$, then

$$\sum_{n=N+1}^{\infty} x_n = \lim_n t_n = \lim_n (s_N + t_n) - s_N = \lim_n s_{N+n} - s_N = x - s_N = \sum_{n=1}^{\infty} x_n - \sum_{n=1}^{N} x_n$$

as required.

(b) If $\sum_n x_n$ converges, then so does every tail series $\sum_{k=n+1}^{\infty} x_k$, by (a). Similarly, if *some* tail series converges, then so does $\sum_n x_n$, and hence so do *all* tail series (again by (a)). Since

$$\sum_k x_k = \sum_{k=1}^{n} x_k + \sum_{k=n+1}^{\infty} x_k = s_n + \sum_{k=n+1}^{\infty} x_k$$

we see that $s_n \to \sum_k x_k$, if and only if $\sum_{k=n+1}^{\infty} x_k \to 0$.

⊣

2.8.2 Absolute Convergence and Rearrangement of Series

Series are sequences of partial sums, and there is always a strong temptation to manipulate them in the same way as finite sums. That can be dangerous, as the following example makes clear:

Example 2.8.9 We have seen that the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges, but that the alternating harmonic series $\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n}$ converges. Suppose that this latter series converges to a number $s \in \mathbb{R}$, i.e.

$$s = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \ldots$$

and let s_n be the n^{th} partial sum.

Now *rearrange the terms* of the alternating harmonic series to obtain a series

$$1-\frac{1}{2}-\frac{1}{4}+\frac{1}{3}-\frac{1}{6}-\frac{1}{8}+\frac{1}{5}-\frac{1}{10}-\frac{1}{12}+\dots$$

and let t_n be the n^{th} partial sum of the rearranged series. Then

$$\begin{aligned}
t_{3n} &= 1-\frac{1}{2}-\frac{1}{4}+\frac{1}{3}-\frac{1}{6}-\frac{1}{8}+\cdots+\frac{1}{2n-1}-\frac{1}{4n-2}-\frac{1}{4n}\\
&= \Big(1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\cdots+\frac{1}{2n-1}\Big)-\Big(\frac{1}{2}+\frac{1}{6}+\frac{1}{10}+\cdots+\frac{1}{4n-2}\Big)\\
&\quad-\Big(\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{4n}\Big)\\
&= \Big(1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\cdots+\frac{1}{2n-1}\Big)-\frac{1}{2}\Big(1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\cdots+\frac{1}{2n-1}\Big)\\
&\quad-\frac{1}{2}\Big(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2n}\Big)\\
&= \frac{1}{2}\Big(1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots+\frac{1}{2n-1}-\frac{1}{2n}\Big)\\
&= \frac{1}{2}s_{2n}
\end{aligned}$$

Thus $\lim_n t_{3n} = \lim_n \frac{1}{2}s_{2n} = \frac{1}{2}s$. It follows easily that $t_n \to \frac{1}{2}s$. Thus the *rearranged* series converges, but not to the same limit as the original.

□

As we shall see, the "problem" is that the harmonic series diverges, but that the alternating harmonic series does not. We therefore define:

Definition 2.8.10 A series $\sum_k x_k$ is said to be *absolutely convergent* if and only if the series of absolute values $\sum_k |x_k|$ is convergent.
A series which is convergent but not absolutely convergent is called *nonabsolutely* or *conditionally convergent.*

Of course, a series of non–negative terms is convergent if and only if it is absolutely convergent.

The notion of absolute convergence is stronger than that of convergence:

Proposition 2.8.11 *An absolutely convergent sequence is convergent.*

Exercise 2.8.12 (a) Prove Proposition 2.8.11.
[Hint: Use the Cauchy criterion, combined with the fact that $\big|\sum_{k=n+1}^{m} x_k\big| \le \sum_{k=n+1}^{m} |x_k|$.]

(b) Exhibit a counterexample that shows that not every convergent series is absolutely convergent.
[Hint: Example 2.8.9]

□

Remarks 2.8.13 (a) From the discussion so far, it should be clear that the series $\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^p}$ is (i) divergent if $p<1$, (ii) conditionally convergent if $p=1$, and (iii) absolutely convergent if $p>1$.

(b) Suppose that $\sum_n x_n$ is conditionally convergent. Then the sequence $\langle x_n \rangle_n$ must change signs infinitely often, i.e. infinitely many x_n are positive, and infinitely are negative.

To see this: Assume, for the sake of argument, that only finitely many x_n are positive, i.e. $\langle x_n \rangle_n$ is not positive infinitely often. Then it must be non–positive eventually[3]. It follows that from some N onwards, the x_n are negative (or zero). Thus there is N such that $|x_n| = -x_n$ for all $n > N$ (because $|a| = -a$ if $a \leq 0$).

Now because $\sum_n x_n$ converges, so does $\sum_{n=N+1}^{\infty} x_n$ (by Proposition 2.8.8), and thus so does $\sum_{n=N+1}^{\infty} |x_n|$ (because $x_n = -|x_n|$ for $n > N$). A tail of the series $\sum_{n=1}^{\infty} |x_n|$ therefore converges, which immediately implies that $\sum_{n=1}^{\infty} |x_n|$ itself converges as well (again by Proposition 2.8.8), i.e. that $\sum_n x_n$ is absolutely convergent — contradiction.

A similar argument holds for the case where only finitely many of the x_n are negative.

□

We now tackle the problem of rearrangements of series which are absolutely convergent. As we shall see, the problem disappears. But first, we need a definition of *rearrangement*:

Definition 2.8.14 Let $f : \mathbb{N} \to \mathbb{N}$ be a bijection. The series $\sum_{n=1}^{\infty} x_{f(n)}$ is called a *rearrangement* of a series $\sum_{k=1}^{\infty} x_k$.

To say that $f : \mathbb{N} \to \mathbb{N}$ is a bijection is equivalent to saying that the list

$$f(1), f(2), f(3), \ldots$$

contains every member of $\mathbb{N}$ once and only once. Essentially, a bijection of $\mathbb{N}$ to $\mathbb{N}$ gives us the elements of $\mathbb{N}$ in rearranged order. The rearranged series

$$x_3 + x_7 + x_1 + x_9 + x_5 + \ldots$$

is obtained from $\sum_{n=1}^{\infty} x_n$ by a bijection f having

$$f(1) = 3, f(2) = 7, f(3) = 1, f(4) = 9, f(5) = 5, \ldots$$

You can easily check that the rearranged alternating harmonic series of Example 2.8.9 is obtained from the alternating harmonic series by the bijection

$$f(3n-2) = 2n-1, f(3n-2) = 4n-2, f(3n) = 4n$$

We are now able to formulate the main result of this section. The proof is left as an exercise:

Theorem 2.8.15 *If $\sum_{n=1}^{\infty} x_k$ converges absolutely, then every rearrangement of $\sum_{n=1}^{\infty} x_k$ converges, and to the same limit.*

Exercise 2.8.16 We prove Theorem 2.8.15.
Suppose that $\sum_{n=1}^{\infty} x_n = s$, and let $\langle s_n \rangle_n$ be the associated partial sums. Let $f : \mathbb{N} \to \mathbb{N}$ be a bijection, and let $x'_n = x_{f(n)}$. We must show that for the rearranged series, we have $\sum_{n=1}^{\infty} x'_n = s$ as well. Let $\langle s'_n \rangle_n$ be the partial sums of the rearranged series.

[3]Because $\neg(\langle x_n \rangle_n$ has P i.o.$) \equiv (\langle x_n \rangle_n$ has $\neg P$ ev.). cf. Remarks 2.2.3

(a) Let $\varepsilon > 0$. Explain why we may choose $N \in \mathbb{N}$ such that $m > n \geq N$ implies $\sum_{k=n+1}^{m} |x_k| < \varepsilon$.

(b) Explain why we may choose $N' \in \mathbb{N}$ such that

$$\{1, 2, \ldots, N\} \subseteq \{f(1), f(2), \ldots, f(N')\}$$

[Hint: Consider $\max\{f^{-1}(1), f^{-1}(2), \ldots, f^{-1}(N)\}$]

(c) Explain why the terms $x_1, x_2, \ldots, x_N$ occur in both s_n and s'_n if $n > N'$. These terms will therefore cancel in the expression $s'_n - s_n$.

(d) It follows that if $n > N'$, then $s'_n - s_n$ contains only terms x_k for which $k > N$. Explain why we must then have

$$|s'_n - s_n| < \varepsilon \qquad \text{whenever} \qquad n > N'$$

(e) Finally, note that

$$|s'_n - s| \leq |s'_n - s_n| + |s_n - s|$$

and conclude that $s'_n \to s$

□

Remarks 2.8.17 Here's an interesting fact: Suppose that $\sum_n x_n$ is a conditionally convergent series, i.e. that it converges, but not absolutely. Let c be any real number whatsoever. Then we can find a rearrangement of $\sum_n x_n$ which converges to c.

The proof of this assertion is not hard, and we will merely outline it here: Given a number $c > 0$, say, add successive non–negative terms of x_n until they exceed c. (This is possible by the next exercise, Exercise 2.8.18). Then add successive negative terms until we obtain a partial sum less than c, then add unused successive positive terms until the partial sum exceeds c, then unused successive negative terms until.... Since $\sum_n x_n$ converges, the terms $x_n \to 0$, so it is not hard to see that this process will result in a rearrangement of $\sum_n x_n$ that converges to c.

□

The next exercise shows that the positive and negative terms of a conditionally convergent series must "add up" to $\pm\infty$:

Exercise 2.8.18 Suppose that $\sum_n x_n$ is conditionally convergent. Let $\langle y_n \rangle_n, \langle z_n \rangle_n$ be the subsequences of $\langle x_n \rangle_n$ consisting, respectively, of non–negative and negative terms. We show that $\sum_n y_n = +\infty$, $\sum_n z_n = -\infty$.

Suppose that $\sum_n x_n = x$. Let X_n, Y_n, Z_n, be, respectively the n^{th} partial sums of $\sum_n x_n, \sum_n y_n$ and $\sum_n z_n$. Furthermore, let A_n be the n^{th} partial sum of $\sum_n |x_n|$.

(a) Explain why $X_n = Y_n + Z_n$, and why $A_n = Y_n - Z_n$.

(b) Explain why $\langle Y_n \rangle_n$ converges if and only if it is bounded. Derive a similar result for $\langle Z_n \rangle_n$.

(c) Suppose now that $\sum_n y_n \neq \infty$. Explain why $\sum_n y_n$ converges.

(d) Let $\sum_n y_n = y$ be the limit. Note that $Z_n = X_n - Y_n$. Explain why $\langle Z_n >_n$ converges.

(e) Conclude that $\langle A_n \rangle_n$ converges, and thus that $\sum_n |x_n|$ converges.

(f) Make sure that you understand why we have now obtained a contradiction from the assumption that $\sum_n y_n \neq \infty$, and that you are able to provide a similar contradiction from the assumption $\sum_n z_n \neq -\infty$

□

2.8.3 More Tests for Convergence

In previous sections, we have obtained several results which guarantee convergence, e..g. Theorems 2.8.1, 2.8.3 and 2.8.5. We derive here a few more results of that ilk.

Theorem 2.8.19 (Comparison Test)
Suppose that $\sum_n x_n$ is a convergent series and that $|y_n| \leq x_n$ for all $n \in \mathbb{N}$ (or merely eventually*). Then $\sum_n y_n$ converges absolutely.*

Proof: Let s_n be the n^{th} partial sum of $\sum_n |y_n|$. We shall show that $\langle s_n \rangle_n$ is a Cauchy sequence. So let $\varepsilon > 0$. Since $\sum_n x_n$ converges, the sequence of tail series $\langle \sum_{k=n+1}^{\infty} x_k \rangle_n$ must converge to 0 (by Proposition 2.8.8), and thus there is an N such that $\sum_{n=N+1}^{\infty} x_k < \varepsilon$.

Then if $n > m > N$, we have

$$\begin{aligned} |s_n - s_m| &= |y_{m+1} + y_{m+2} + \cdots + y_n| \\ &\leq |y_{m+1}| + |y_{m+2}| + \cdots + |y_n| \\ &\leq x_{m+1} + x_{N+2} + \cdots + x_n \\ &\leq \sum_{k=N+1}^{\infty} x_k < \varepsilon \end{aligned}$$

Hence $\langle s_n \rangle_n$ is a Cauchy sequence, and thus convergent.

A minor modification, invoking Proposition 2.8.8, shows that the result remains true if we only have $|y_n| \leq x_n$ eventually.

⊣

Theorem 2.8.20 (Root Test)
Suppose that $\sum_n x_n$ is a series in $\mathbb{R}$:

(a) If there is $\alpha < 1$ such that $|x_n|^{\frac{1}{n}} \leq \alpha$ eventually, then $\sum_n x_n$ is absolutely convergent.

(b) If $|x_n|^{\frac{1}{n}} \geq 1$ infinitely often, then $\sum_n x_n$ diverges.

Proof: (a) There is $N \in \mathbb{N}$ such that

$$|x_n| < \alpha^n \quad \text{whenever} \quad n > N$$

Now since $\alpha < 1$, the geometric series $\sum_n \alpha^n$ converges. By the Comparison Test, the series $\sum_n |x_n|$ converges as well.

(b) If $|x_n|^{\frac{1}{n}} \geq 1$ infinitely often, then $|x_n| \geq 1$ infinitely often, so certainly $x_n \not\to 0$. Thus $\sum_n x_n$ cannot be convergent (by Corollary 2.8.2).

⊣

Remarks 2.8.21 (1) Suppose that $\lim_n |x_n|^{\frac{1}{n}} =: \alpha$ exists. If $\alpha < 1$ then $\sum_n x_n$ cponverges absolutely. If $\alpha > 1$, then $\sum_n x_n$ diverges.

(2) If $\lim_n |x_n|^{\frac{1}{n}} = 1$, then the Root Test is inconclusive. Recall that $n^{\frac{1}{n}} \to 1$ — see Exercise 2.3.8.
If $x_n = \frac{1}{n}, y_n = \frac{1}{n^2}$, then $\limsup_n |x_n|^{\frac{1}{n}} = 1$ and also $\limsup_n |y_n|^{\frac{1}{n}} = 1$. Now $\sum_n x_n$ diverges, whereas $\sum_n y_n$ converges absolutely. It follows that the Root Test is inconclusive if $\alpha = 1$.

□

Theorem 2.8.22 (Ratio Test)
Suppose that $\sum_n x_n$ is a series with $\lim_n \left|\frac{x_{n+1}}{x_n}\right| = \alpha$. If $\alpha < 1$, then $\sum_n x_n$ converges absolutely, and if $\alpha > 1$, then $\sum_n x_n$ diverges.
If $\alpha = 1$, the test is inconclusive.

Proof: Suppose first that $\alpha < 1$. Choose β such that $\alpha < \beta < 1$. Then $\left|\frac{x_{n+1}}{x_n}\right| < \beta$ eventually, i.e. there is $N \in \mathbb{N}$ such that

$$\left|\frac{x_{n+1}}{x_n}\right| < \beta \quad \text{whenever} \quad n \geq N$$

Then

$$\begin{aligned} |x_{N+1}| &< \beta |x_N| \\ |x_{N+2}| &< \beta |x_{N+1}| < \beta^2 |x_N| \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ |x_{N+n}| &\leq \beta^n |x_N| \end{aligned}$$

Now $\sum_n \beta^n |x_N| = |x_N| \sum_n \beta^n$ converges, and thus by the Comparison Test, the series $\sum_n |x_n|$ converges as well.

If $\alpha > 1$, then $|x_{n+1}| \geq |x_n|$ eventually, so we cannot have $x_n \to 0$. Thus $\sum_n x_n$ diverges.

⊣

Examples 2.8.23 Let $x \in \mathbb{R}$. We show that the series

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}$$

converges. We apply the Ratio Test. The ratio of successive terms is

$$\frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}} = \frac{x}{n+1}$$

Now, no matter how big x is, $\frac{x}{n+1} \to 0$ as $n \to \infty$. Since $0 < 1$, the Ratio Test guarantees that $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ converges absolutely.
And you probably already know that it converges to e^x.

□

Exercise 2.8.24 Determine whether or not the following series converge.

(a) $\sum_{n=1}^{\infty} \frac{n+1}{2n-1}$

(b) $\sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!}$

(c) $\sum_{n=1}^{\infty} (-1)^{n-1} n^{-\frac{1}{3}}$

(d) $\sum_{n=1}^{\infty} \frac{\ln n}{n^2}$

(e) $\sum_{n=2}^{\infty} \frac{1}{n \ln n}$

(f) $\sum_{n=1}^{\infty} \frac{\sqrt{n+1}-\sqrt{n}}{n}$

(g) $\sum_{n=1}^{\infty} n^{-\sqrt{n}}$

(h) $\sum_{n=1}^{\infty} \frac{1}{(\ln n)^n}$

□

Exercise 2.8.25 (a) Suppose that $\sum_n x_n$ and $\sum_n y_n$ are series with non–negative terms, and that $\frac{x_n}{y_n} \to l$ as $n \to \infty$, where $l \neq 0$. Prove that $\sum_n x_n$ converges if and only if $\sum_n y_n$ converges.

(b) Hence determine whether or not the following sequences converge:

$$(i) \ \sum_n \frac{1}{2n-1} \qquad (ii) \ \sum_n \frac{5}{2n^2+4}$$

[Hint:(a) If $\frac{x_n}{y_n} \to l$, then eventually $y_n(l+\varepsilon) > x_n$. It follows that there is $C > 0$ such that $x_n \leq C y_n$ for all n. Now use Comparison.]

□

Here is another result on convergence of series which is slightly ahead of time: Recall (from first–year calculus) that if f is continuous on the interval $[a, \infty)$, then the *improper integral* $\int_a^{\infty} f(x)\, dx$ is defined by

$$\int_a^{\infty} f(x)\, dx = \lim_{b \to \infty} \int_a^b f(x)\, dx$$

whenever this limit exists. If this limit exists, the improper integral is said to *converge*. (We are ahead of time because we have not yet said what we mean by $\lim_{b\to\infty}$. At this stage, you are no doubt able to provide the definition yourself, however.)

Proposition 2.8.26 (Integral Test for Convergence) *Suppose that f is continuous, decreasing and non–negative on the interval $[1, \infty)$. Further supposes that $\sum_{n=1}^{\infty} y_n$ is a series with $y_n = f(n)$ for $n \in \mathbb{N}$. Then the improper integral $\int_1^{\infty} f(x)\, dx$ converges if and only if the series $\sum_n y_n$ converges.*

Exercise 2.8.27 We prove Proposition 2.8.26:

(a) Explain why $y_{k+1} \leq f(x) \leq y_k$ whenever $k \leq x \leq k+1$ (where $k \in \mathbb{N}$)

(b) Show that $y_{k+1} \leq \int_k^{k+1} f(x)\,dx \leq y_k$ for all $k \in \mathbb{N}$.

(c) Hence show that

$$\sum_{k=2}^{n} y_k \leq \int_1^n f(x)\,dx \leq \sum_{k=1}^{n-1} y_k$$

for all $n \in \mathbb{N}$.

(d) Now suppose that $\int_1^\infty f(x)\,dx$ converges, i.e. that $\lim_{b\to\infty} \int_1^b f(x)\,dx$ exists and is $< \infty$. Explain why $\sum_{k=1}^\infty y_k$ converges as well.

(e) Next, assume that $\int_1^\infty f(x)\,dx$ diverges. Show that $\sum_{n=1}^\infty y_n$ diverges as well.

(f) The proof of the proposition is now complete. Now use this result to determine whether or not the series

$$\sum_{n=3}^{\infty} \frac{1}{n \ln n}$$

converges. [Hint: Note that $\sum_{n=3}^\infty \frac{1}{n\ln n} = \sum_{n=1}^\infty \frac{1}{(n+2)\ln(n+2)}$]

□

2.9 lim sup and lim inf*

Suppose that $\langle x_n\rangle_n$ is a *bounded* sequence in $\mathbb{R}$. Construct two new sequences as follows:

$$y_n = \sup\{x_m : m \geq n\} \qquad z_n = \inf\{x_m : m \geq n\}$$

Because $\langle x_n\rangle$ is bounded, y_n and z_n exist (i.e. are finite real numbers), by the Completeness Axiom.

Suppose, for example, that $x_n = \frac{(-1)^n}{n}$ for $n \geq 1$. Then

$$y_1 = \sup\left\{-1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\right\} = \frac{1}{2}$$
$$y_2 = \sup\left\{\frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\right\} = \frac{1}{2}$$
$$y_3 = \sup\left\{-\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \dots\right\} = \frac{1}{4}$$
$$y_4 = \sup\left\{\frac{1}{4}, -\frac{1}{5}, \dots\right\} = \frac{1}{4}$$

i.e. $\langle y_n\rangle$ is the sequence $\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{8}, \dots$.

Similarly, you can check that $\langle z_n\rangle$ is the sequence $-1, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{7}, -\frac{1}{7}, -\frac{1}{9}\dots$.

Exercise 2.9.1 (a) Given $\langle x_n\rangle_n$, write down the first 6 terms of $y_n = \sup\{x_m : m \geq n\}$ and $z_n = \inf\{x_m : m \geq n\}$.

(i) $x_n = (-1)^n$

(ii) $x_n = \frac{1}{n}$

(iii) $x_n = \begin{cases} 1 + \dfrac{1}{n} & \text{if } n \text{ is odd} \\ -1 - 2^{-n} & \text{if } n \text{ is even} \end{cases}$

(iv) $x_n = \begin{cases} 1 - \dfrac{1}{n} & \text{if } n \text{ is odd} \\ -1 + 2^{-n} & \text{if } n \text{ is even} \end{cases}$

(b) Note that each of the above sequences $\langle y_n \rangle$ is decreasing, and that each of the $\langle z_n \rangle_n$ is increasing. Can you explain why?

(c) Finally, since the $\langle y_n \rangle$ and $\langle z_n \rangle$ are bounded monotone sequences, they must converge (by Theorem 2.3.11). Write down $\lim_n y_n$ and $\lim_n z_n$ for each of the sequences in (a)(i)-(iv).

□

As noted in the above exercise, $\langle y_n \rangle$ is a decreasing sequence, and $\langle z_n \rangle$ is increasing. To see this, let $A_n = \{x_m : m \geq n\}$. Clearly

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$$

Hence

$$\sup A_1 \geq \sup A_2 \geq \sup A_3 \geq \dots \qquad \text{and} \qquad \inf A_1 \leq \inf A_2 \leq \inf A_3 \leq \dots$$

(Note that if $A \subseteq B$, then $\sup A \leq \sup B$, and $\inf A \geq \inf B$.)
Since $y_n = \sup A_n$ and $z_n = \inf A_n$, and because $\sup A_n \geq \inf A_n$, we see that

$$z_1 \leq z_2 \leq z_3 \leq \dots \leq z_n \leq \dots \dots \leq y_n \leq \dots \leq y_3 \leq y_2 \leq y_1$$

Now any bounded monotone sequence converges (Theorem 2.3.11), and thus $\lim_n y_n$ and $\lim_n z_n$ exist if $\langle x_n \rangle$ is bounded. We now define $\limsup_n x_n = \lim_n y_n$, and $\liminf_n x_n = \lim_n z_n$:

Definition 2.9.2 Let $\langle x_n \rangle$ be a sequence in $\mathbb{R}$. We define the *limit superior* of $\langle x_n \rangle$ by

$$\limsup_n x_n = \lim_{n \to \infty} \sup_{m \geq n} x_m$$

where we adopt the convention that if $\langle x_n \rangle$ is not bounded above, we set $\limsup_n x_n = +\infty$.
Similarly, we define the *limit inferior* of $\langle x_n \rangle$ by

$$\liminf_n x_n = \lim_{n \to \infty} \inf_{m \geq n} x_m$$

where we adopt the convention that if $\langle x_n \rangle$ is not bounded below, we set $\liminf_n x_n = -\infty$.

The notions of lim sup and lim inf are quite difficult, so we will approach them in another way. Let $\langle x_n \rangle_n$ be a sequence, and let $y_n := \sup_{m \geq n} x_m$ and $z_n := \inf_{m \geq n} x_m$.

- If $\limsup_n x_n > a$, then $\lim y_n > a$. Hence $y_n > a$ for all n (since $\langle y_n \rangle_n$ is decreasing). It follows that, for all n, $\sup_{m \geq n} x_m > a$, i.e. that there exists $m \geq n$ such that $x_m > a$. Thus

$$\limsup_n x_n > a \qquad \Longrightarrow \qquad \forall n \in \mathbb{N}\ \exists m \geq n (x_m > a)$$

i.e.

$$\limsup_n x_n > a \qquad \Longrightarrow \qquad x_n > a \text{ infinitely often}$$

- On the other hand, if $x_n \geq a$ infinitely often, then $y_n := \sup_{m \geq n} x_m \geq a$ for all n. Thus $\limsup_n x_n = \lim_n y_n \geq a$, i.e.

$$x_n \geq a \text{ infinitely often} \qquad \Longrightarrow \qquad \limsup_n x_n \geq a$$

- From the logical equivalence of $\varphi \to \psi$ and $\neg\psi \to \neg\varphi$, we see that

$$x_n \leq a \text{ eventually} \qquad \Longrightarrow \qquad \limsup_n x_n \leq a$$

and

$$\limsup_n x_n < a \qquad \Longrightarrow \qquad x_n < a \text{ eventually}$$

- Since $\liminf_n x_n = \sup_n \inf_{m \geq n} x_m = -\inf_n \sup_{m \geq n}(-x_m) = -\limsup_n(-x_n)$ (because $-\sup A = \inf(-A)$, where $-A := \{-a : a \in A\}$), we need to prove a result only for lim sup, in order to get immediately a corresponding result for lim inf. Similar statements therefore hold for lim inf.

Summarizing in a box:

$$\begin{array}{rcl}
\limsup_n x_n > a & \Longrightarrow & x_n > a \text{ infinitely often} \\
x_n \geq a \text{ infinitely often} & \Longrightarrow & \limsup_n x_n \geq a \\
x_n \leq a \text{ eventually} & \Longrightarrow & \limsup_n x_n \leq a \\
\limsup_n x_n < a & \Longrightarrow & x_n < a \text{ eventually} \\
\liminf_n x_n < a & \Longrightarrow & x_n < a \text{ infinitely often} \\
x_n \leq a \text{ infinitely often} & \Longrightarrow & \liminf_n x_n \leq a \\
x_n \geq a \text{ eventually} & \Longrightarrow & \liminf_n x_n \geq a \\
\liminf_n x_n > a & \Longrightarrow & x_n > a \text{ eventually}
\end{array}$$

If you understand the implications in the box, you understand lim sup and lim inf.

Proposition 2.9.3 *Suppose that $\langle x_n \rangle_n$, $\langle y_n \rangle_n$ are bounded sequences in $\mathbb{R}$, and that $\lambda \in \mathbb{R}$.*

(a) $\liminf_n x_n \leq \limsup_n x_n$

(b) If $\lambda \geq 0$, then $\limsup_n \lambda x_n = \lambda \limsup_n x_n$, *and* $\liminf_n \lambda x_n = \lambda \liminf_n x_n$

(c) If $\lambda < 0$, then $\limsup_n \lambda x_n = \lambda \liminf_n x_n$, *and* $\liminf_n \lambda x_n = \lambda \limsup_n x_n$

(d) $\limsup_n (x_n + y_n) \leq \limsup_n x_n + \limsup_n y_n$

(e) $\liminf_(x_n + y_n) \geq \liminf_n x_n + \liminf_n y_n$

(f) If $x_n \leq y_n$, then $\limsup_n x_n \leq \limsup_n y_n$ *and* $\liminf_n x_n \leq \liminf_n y_n$

Proof: Here's a proof of (a): Suppose that $x := \liminf_n x_n$, and that $\varepsilon > 0$. Then $\liminf_n x_n > x - \varepsilon$, so $x_n > x - \varepsilon$ eventually. Obviously then also $x_n \geq x - \varepsilon$ infinitely often, so that It follows that $\limsup_n x_n \geq x - \varepsilon$. Since this holds for all $\varepsilon > 0$ we must [4] have $\limsup_n x_n \geq x$, as required

The rest of this proposition is left as an exercise.

⊣

Exercise 2.9.4 Prove the remainder of Proposition 2.9.3.
[Hints: (c) If $x_n > z$ infinitely often and $\lambda < 0$, then $\lambda x_n < \lambda z$ infinitely often.
(d) If $z > \limsup_n x_n$ and $w > \limsup_n y_n$, then $x_n < z$ eventually and $y_n < w$ eventually. Hence $x_n + y_n < z + w$ eventually.]

□

Though a bounded sequence $\langle x_n \rangle_n$ may not have a limit, it always has a lim sup and a lim inf. When $\langle x_n \rangle_n$ *does* converge, the three notions coincide, and conversely, as we shall see next. Note that always $\limsup_n x_n \geq \liminf_n x_n$:

Proposition 2.9.5 *Suppose that $\langle x_n \rangle_n$ is a bounded sequence of real numbers. Then $\langle x_n \rangle_n$ converges if and only if $\limsup_n x_n = \liminf_n x_n$. In that case, $\lim_n x_n = \limsup_n x_n = \liminf_n x_n$.*

Proof: ($\Rightarrow$): Suppose that $x_n \to x$, and let $\varepsilon > 0$. Then $|x_n - x| < \varepsilon$ eventually, and thus in particular $x_n \leq x + \varepsilon$ eventually. Thus $\limsup_n x_n \leq x + \varepsilon$.

Similarly $x - \varepsilon \leq x_n$ eventually, and thus $\liminf_n x_n \geq x - \varepsilon$.

It follows that for all $\varepsilon > 0$, we have

$$x - \varepsilon \leq \liminf_n x_n \leq \limsup_n x_n \leq x + \varepsilon$$

Since ε was arbitrary, we must have $\liminf_n x_n = \limsup_n x_n = x$.

($\Leftarrow$): Suppose that $\liminf_n x_n = \limsup_n x_n =: x$, and let $\varepsilon > 0$ be arbitrary. Then $\liminf_n x_n > x - \varepsilon$, so $x_n > x - \varepsilon$ eventually. Similarly, $\limsup_n x_n < x + \varepsilon$, so $x_n < x + \varepsilon$ eventually. Combining, we see that $x - \varepsilon < x_n < x + \varepsilon$ eventually, i.e. that $|x_n - x| < \varepsilon$ eventually.

⊣

Proposition 2.9.6 *Every sequence of real numbers has a monotone subsequence.*
In fact every sequence has a monotone subsequence which converges to $\limsup_n x_n$ (and similarly, one which converges to $\liminf_n x_n$).

Proof: Given a sequence $\langle x_n \rangle_n$, we show that there is a monotone subsequence which converges to $\limsup_n x_n$. Let $\bar{x} = \limsup_n x_n$, and put $y_n = \sup_{m \geq n} x_m$ (so that $y_n \downarrow \bar{x}$). We distinguish two cases.

Case 1: $\bar{x} < y_n$ for all n.
(We allow here the case $\bar{x} = -\infty$.) In that case, we can choose a *decreasing* subsequence

[4] What we are using here is that, if $a \geq x - \varepsilon$ for all $\varepsilon > 0$, then also $a \geq x$. For if not, then $a < x$. But if we now define $\varepsilon := \frac{x-a}{2}$, then $\varepsilon > 0$ and $a < x - \varepsilon$ — contradiction.

$\langle x_{n_k}\rangle_k$ inductively, as follows: Let $N_1 = 1$. Since $y_{N_1} > \bar{x}$, there is $n_1 \geq N_1$ so that $x_{n_1} > \bar{x}$. Next, since $y_n \downarrow \bar{x}$, there is $N_2 > n_1$ such that $y_{N_2} < x_{n_1}$. Since, by hypothesis $y_{N_2} > \bar{x}$, there is $n_2 \geq N_2$ such that $x_{n_2} > \bar{x}$ also. Thus $\bar{x} < x_{n_2} \leq y_{N_2} < x_{n_1} \leq y_{N_1}$.

Keep going in the same way: Once we have constructed a double sequence of integers $N_1 \leq n_1 < N_2 \leq n_2 < \cdots < N_k \leq n_k$ such that

$$x_{n_1} > x_{n_2} > \cdots > x_{n_{k-1}} > x_{n_k} > \bar{x}$$

we may choose $N_{k+1} > n_k$ so that $y_{N_{k+1}} < x_{n_k}$. Since also $y_{N_{k+1}} > \bar{x}$, there is $n_{k+1} \geq N_{k+1}$ such that $x_{n_{k+1}} > \bar{x}$. Thus $n_{k+1} \geq N_{k+1} > n_k$ and $x_{n_k} > y_{N_{k+1}} \geq x_{n_{k+1}} > \bar{x}$.

This completes the inductive construction of the subsequence $\langle x_{n_k}\rangle_k$. Now $y_n \to \bar{x}$, and so also the subsequence $y_{N_k} \to \bar{x}$, by Proposition 2.6.4. Since $\bar{x} < x_{n_k} \leq y_{N_k}$ for all k, the Sandwich Theorem ensures that $x_{n_k} \to \bar{x}$ as well.

Case 2: There is N_0 such that $y_{N_0} = \bar{x}$.
(We allow here the case $\bar{x} = +\infty$.) In that case, since y_n is a decreasing sequence converging to $\bar{x}$, we must have $y_n = \bar{x}$ for all $n \geq N_0$ also. In particular, it follows that $x_n \leq \bar{x}$ for all $n \geq N_0$. Thus *either* (i) $x_n = \bar{x}$ infinitely often, *or* (ii) $x_n < \bar{x}$ eventually. If (i) holds, there is obviously a constant (hence monotone) subsequence converging to $\bar{x}$, so it remains to deal with (ii).

Suppose therefore that $x_n < \bar{x}$ for all $n \geq N_1$, and let $N = \max\{N_0, N_1\}$. Then

$$\forall n \geq N \ (y_n = \bar{x} \wedge x_n < \bar{x})$$

Define $t_n = \bar{x} - \frac{1}{n}$ if $\bar{x}$ is finite, and put $t_n = n$ if $\bar{x}$ is infinite. Then $t_n < \bar{x}$, and $t_n \uparrow \bar{x}$, whether $\bar{x}$ is finite or not. Inductively construct an increasing subsequence x_{n_k} as follows: Choose $n_1 \geq N$, so that $t_1 \leq x_{n_1} < \bar{x}$. Now $y_{n_1+1} = \bar{x}$, because $y_n = \bar{x}$ for all $n \geq N$, and so there is $n_2 > n_1$ such that $\max\{x_{n_1}, t_2\} < x_{n_2}$. Of course, also $x_{n_2} < \bar{x}$. Proceed in the same way: Given $n_1 < n_2 < \cdots < n_k$ such that

$$\max\{x_{n_j}, t_{j+1}\} < x_{n_{j+1}} < \bar{x} \qquad \text{for } j = 1, \ldots, k-1$$

choose $n_{k+1} > n_k$ such that $x_{n_{k+1}} > \max\{x_{n_k}, t_{k+1}\}$ — this is possible because $y_{n_k+1} = \sup\{x_n : n > n_k\} = \bar{x}$.

In this way we obtain a strictly increasing subsequence $\langle x_{n_k}\rangle_k$ such that $t_k < x_{n_k} < \bar{x}$. Since $t_k \to \bar{x}$, we see that $x_{n_k} \to \bar{x}$ also, by the Sandwich Theorem.

$\dashv$

Chapter 3

Basic Topology

3.1 Introduction

The aim of this short chapter is to create a new *language* for talking about *space*. This language, couched in the terminology of sets, is extremely general, and applies to structures vastly different from $\mathbb{R}$, although we will mainly apply it to the reals. We are going to define a large number of simple concepts, and state a large number of simple propositions. Indeed, *all* of the propositions in this section are trivial, in that they only require one to plug in the appropriate definitions to prove them. But those definitions take *some* getting used to! It is therefore *extremely important* that you do all the exercises — perhaps several times over! There is no other way to learn this new language.

Most of the concepts we define will invoke only the notion of *distance* $d(x, y)$ between two points. For example, in $\mathbb{R}$, the usual distance between x, y is defined by

$$d(x, y) := |x - y| = \sqrt{(x - y)^2}$$

In $\mathbb{R}^n$, the usual distance is given by $d(\mathbf{x}, \mathbf{y}) := \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where $\mathbf{x} := (x_1, x_2, \ldots, x_n)$.

Such a distance function d is called a *metric*, and the properties we require it to have are extremely simple:

Definition 3.1.1 A *metric space* is a pair (X, d) consisting of a set X together with a map $d : X \times X \to \mathbb{R}$, called a *metric*, which satisfies the following conditions:

(i) $d(x, y) \geq 0$ for all $x, y \in X$;

(ii) $d(x, y) = 0$ if and only if $x = y$;

(iii) $d(x, y) = d(y, x)$ for all $x, y \in X$;

(iv) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (Triangle Inequality);

Exercise 3.1.2 Verify that the map $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R} : (x, y) \mapsto |x - y|$ is a metric, i.e. that it satsfies (i)-(iv) in Defn. 3.1.1. Note that (iv) is equivalent to the usual triangle inequality in $\mathbb{R}$. The triangle inequality has already been used many times in these notes to provide important estimates. It plays a similar role in the theory of metric spaces.

□

Remarks 3.1.3 Many of the notions that we have studied so far carry over immediately to arbitrary metric spaces. For example:

- Let $\langle x_n \rangle$ be a sequence in a metric space (X, d). Then $x_n \to x$ in the space X should mean that the distance between x_n and x converges to 0, i.e. that $d(x_n, x) \to 0$. Since $\langle d(x_n, x) \rangle_n$ is a sequence of (non–negative) real numbers, we have already defined what it means to say $d(x_n, x) \to 0$, and thus we see that

$$x_n \to x \quad \text{in } (X, d) \qquad \Longleftrightarrow \qquad \forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n \geq N \; [d(x_n, x) < \varepsilon]$$

  This is exactly the definition of convergence in $\mathbb{R}$ when d is the usual metric $d(x, y) := |x - y|$.

- As it stands, the Completeness Axiom does not make sense in arbitrary metric spaces, as it depends on the order relation on $\mathbb{R}$. But we saw that one of the most important consequences of the Completeness Axiom in $\mathbb{R}$ is that Cauchy sequences converge. Now we *can* define the notion of Cauchy sequence in an abstract metric space (X, d):

$$\langle x_n \rangle \text{ is a Cauchy sequence} \qquad \Longleftrightarrow \qquad \forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n, m \geq N \; [d(x_n, x_m) < \varepsilon]$$

  We then define a metric space to be *complete* if and only every Cauchy sequence converges. We thus have a definition of completeness that is phrased in terms of distance, rather than order.

□

We will not study general metric spaces in this course, but restrict our attention to $\mathbb{R}$, with its usual metric. We will, however, phrase most of our definitions in terms of the metric, to emphasize the geometric flavour of our definitions. As a bonus, these definitions, as well as many proofs, carry over verbatim to more general spaces.

Remarks 3.1.4 Furthermore, we shall have occasion to use it in a more abstract setting at least once, when we discuss uniform convergence of functions. Here the underlying space is not $\mathbb{R}$, but the set of real–valued continuous functions $C[a, b]$ defined on an interval $[a, b]$. The functions space $C[a, b]$ comes equipped with a metric:

$$d(f, g) := \sup_{x \in [a,b]} |f(x) - g(x)|$$

We say that a sequence of such functions $\langle f_n \rangle$ converges *uniformly* to a function f if $d(f_n, f) \to 0$. This notion is extremely important for proving regularity properties of power series expansions of functions, for example. Unraveling the definition, this means

$$\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \geq N \forall x \in [a, b] \; [|f_n(x) - f(x)| < \varepsilon]$$

This would be the definition of uniform convergence if we don't use the notion of metric, and is rather more complicated.

□

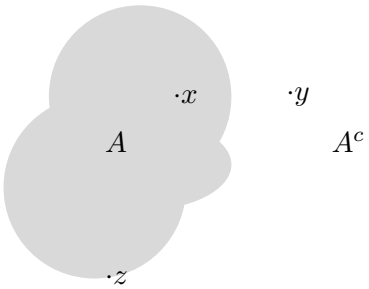
3.2 Open and Closed Sets — Motivation

We already know what we mean by *open interval* and *closed interval*. We would like now to extend the definition of *open* and *closed* to more general subsets of $\mathbb{R}$, and not just intervals. Intuitively, a subset $U \subseteq \mathbb{R}$ is *open* if it contains *none* of its boundary points, and *closed* if it contains *all* of its boundary points. For example, the interval (a, b) has boundary points a, b, and these do not belong to it, so (a, b) is open. On the other hand, the interval $[a, b]$ has the same boundary points, and they belong to it, so $[a, b]$ is closed. An interval such as $(a, b]$ is neither an open set nor a closed set.

In the same way the set $(1, 2) \cup (3, \infty)$ is open, whereas the set $(0, 1)^c$ is closed.

This is all very well for simple sets such as intervals, where we can see the boundary points. But it won't wash for more complicated sets, because we have no definition of *boundary point*. What, for example, are the boundary points of the set $\mathbb{Q}$, or of the singleton set $\{0\}$? Clearly, a deeper analysis of these notions is needed.

For this, it helps to consider the space (X, d) to be $\mathbb{R}^2$ with its usual metric, because then we can visualize the concepts more easily by drawing pictures. Suppose, therefore that A is a non–empty subset of X.



We see here three points x, y, z. Using some standard notions in English (which have not yet been defined mathematically), we see that:

- x is clearly "inside" A. We say that x is in the *interior* of A.
- y is on the "outside" of A. This means that y is on the "inside" of A^c, i.e. y is in the interior of A^c.
- z is on the boundary of A. It is also on the boundary of A^c. It is neither in the interior of A nor in the interior of A^c.

It is thus clear that if we can mathematically define what it means for a point to be "inside" a set A, then we can also define what it means for a point to be on the "outside" and on the boundary. We thus seek a mathematical definition for the notion of an *interior point* of a set.

Now if a point x is truly "inside" a set A, then it should be possible to move a (perhaps very small) distance in *any* direction from x without leaving the set A. This means that there is some (perhaps very small) number $\varepsilon > 0$ such that if you move a distance of $< \varepsilon$ in any direction from x, then you stay inside A. Dispensing with the notion of *direction*, this means that any point which is sufficiently close to the point x also belongs to the set X. Let us now define the *open ball of radius ε centered at x* by:

$$B(x, \varepsilon) := \{x' \in X : d(x, x') < \varepsilon\}$$

If by "sufficiently close" we mean "within a distance of ε" this translates to the following:

$$x \text{ is in the interior of } A \text{ if and only if there is } \varepsilon > 0 \text{ such that } B(x, \varepsilon) \subseteq A$$

Remarks 3.2.1 • Obviously, when we are dealing with $\mathbb{R}$ equipped with its usual metric, an open "ball" is just an open interval:

$$B(x, \varepsilon) = (x - \varepsilon, x + \varepsilon)$$

Conversely an open interval is an open ball:

$$(a, b) = B(\tfrac{b+a}{2}, \tfrac{b-a}{2})$$

- In $\mathbb{R}^2$ an open ball is a circle (not including its boundary), and in $\mathbb{R}^3$ it really is a ball (not including its boundary).
- However, the notion of *open ball* makes sense in any metric space, because its definition uses only the notion of metric, i.e. distance: $B(x, \varepsilon)$ is the set of all points whose distance to the point x is $< \varepsilon$. Hence the definition of *interior point* makes sense in any metric space, as it only uses the notion of open ball, and this only needs the notion of distance.

□

Now that we have mathematically defined the notion of interior point, we can define the notion of boundary point. A point $z \in X$ is a boundary point of A if it is neither an interior point of A nor an interior point of A^c. So let's first analyze what it means to assert that z is not an interior point of a set A: It means that for no $\varepsilon > 0$ is it the case that $B(z, \varepsilon) \subseteq A$. This, in turn, means that for every $\varepsilon > 0$ we must have $B(z, \varepsilon) \cap A^c \neq \emptyset$, i.e. every open ball centered at z must intersect A^c, i.e. must contain a point of A^c. [It may be helpful to recall that $B \subseteq C$ if and only if $B \cap C^c = \emptyset$.]

Since a point $z \in X$ is a boundary point of A if it is neither an interior point of A nor an interior point of A^c, we can now assert that

z is a boundary point of A if and only every open ball centered at z must intersect both A and A^c.

Note that— since $(A^c)^c = A$ — this definition is symmetric in A, A^c. Thus if z is a boundary point of A, then z is also a boundary point of A^c, and vice versa.

Now recall that we want a set $A \subseteq X$ to be *open* if A contains *none* of its boundary points. We want a set $A \subseteq X$ to be *closed* if A contains *all* of its boundary points. If we analyze this, it turns out that we can dispense with the notion of boundary point altogether!

- A is open iff for every $x \in A$ it is the case that x is not a boundary point of A.
  - This means that for every $x \in A$ there is some open ball $B(x, \varepsilon) > 0$ such that either $B(x, \varepsilon) \cap A^c = \emptyset$ or $B(x, \varepsilon) \cap A = \emptyset$.
  - But it is impossible that $B(x, \varepsilon) \cap A = \emptyset$, since $x \in A$ and $x \in B(x, \varepsilon)$.
  - We thus see that for every $x \in A$ there is $\varepsilon > 0$ such that $B(x, \varepsilon) \cap A^c = \emptyset$.

– Hence for every $x \in A$ there is $\varepsilon > 0$ such that $B(x, \varepsilon) \subseteq A$. [Again recall that $B \subseteq C$ if and only if $B \cap C^c = \emptyset$.]

– But this means exactly that every $x \in A$ is an interior point of A (!!)

Thus $A \subseteq X$ is open if and only if every point of A is an interior point of A.

$$A \subseteq X \text{ is open} \qquad \Longleftrightarrow \qquad \forall x \in A \exists \varepsilon > 0 [B(x, \varepsilon) \subseteq A]$$

- $A \subseteq X$ is closed if every boundary point of A belongs to A.

– This means that every boundary point of A^c belongs to A, because A and A^c have the same boundary points.

– Hence no boundary point of A^c belongs to A^c.

– And hence A^c is open (!!).

Thus $A \subseteq X$ is closed if and only if A is open.

We built on the intuition provided by open and closed intervals to define open and closed sets is any metric space. For this, we needed the notion of boundary point. But when we finally analyzed the definitions of open and closed set, it turned out that boundary points were unnecessary — all we need is the notion of interior point. A set is open precisely if all of its points are interior points. A set is closed if its complement is an open set.

□

In the next section we will write down all these definitions again, without the motivation provided here, and proceed to deduce some simple consequences from these definitions.

3.3 Open and Closed Sets — Definitions and Basic Properties

3.3.1 Definitions

Definition 3.3.1 Let (X, d) be a metric space.

(i) For $x_0 \in X$ and $r > 0$, we define the *open ball* of radius r centered at x_0 by

$$B(x_0, r) := \{x \in X : d(x_0, x) < r\}$$

(ii) Let $A \subseteq X$. We say that a point $x \in X$ is an *interior point* of A if and only if there is an open ball centered at x which is contained in A. We denote the set of interior points of A by A°. i.e.

$$x \in A^\circ \qquad \Longleftrightarrow \qquad \exists r > 0[B(x, r) \subseteq A]$$

Note that $A^\circ \subseteq A$.

(iii) We say that $A \subseteq X$ is a *neigbourhood* of $x \in X$ if and only if x is an interior point of A:

$$A \text{ is a neighbourhood of } x \qquad \Longleftrightarrow \qquad x \in A^\circ$$

(iv) We say that $A \subseteq X$ is and *open set* if and only if every element of A is an interior point, if and only if A is a neighbourhood of each of its elements:

$$A \text{ is open} \qquad \Longleftrightarrow \qquad A^\circ = A$$

(v) We say that $A \subseteq X$ is a *closed set* if and only if its complement A^c is open.

(vi) We say that $x \in X$ is a *boundary point* of $A \subseteq X$ if and only if every open ball centered at x has non–empty intersection with both A and A^c, if and only if x is neither an interior point of A^c, nor an interior point of A. We denote the set of boundary points of A by ∂A:

$$x \in \partial A \qquad \Longleftrightarrow \qquad \forall r > 0[B(x, r) \cap A \neq \emptyset \ \wedge B(x, r) \cap A^c \neq \emptyset]$$

Exercise 3.3.2 Let us see what these notions mean in $\mathbb{R}$, i.e. let (X, d) be $\mathbb{R}$ with the usual metric $d(x, y) := |x - y|$:

(a) Show that every open ball is an open interval, and *vice versa*.

(b) Show that every open interval is an open set.

(c) Find an open set which is not an open interval.

(d) Show that every closed interval is a closed set.

(e) Show that $[0, 1]^\circ = (0, 1)$. Deduce that $[0, 1]$ is not an open set.

(f) Show that $(0, 1]$ is neither open, n or closed.

(g) Show that $\mathbb{Q}^\circ = \emptyset$.

(h) Show that $\emptyset$ and $\mathbb{R}$ are both open sets, and also that they are both closed sets.

□

3.3.2 Open Sets

We have $A^\circ \subseteq A$, and thus $\emptyset^\circ = \emptyset$. It follows that $\emptyset$ is always open, in any metric space.

Proposition 3.3.3 *Let (X, d) be a metric space.*
A set is open if and only if it is a (possibly infinite) union of open balls.

Proof: ($\Longrightarrow$): Suppose that $A \subseteq X$ is an open set. If $a \in A$, then a is an interior point of A, and thus there is $r_a > 0$ so that $B(x, r_a) \subseteq A$. It follows easily that $A = \bigcup_{a \in A} B(a, r_a)$, and thus that A is a union of open balls.
($\Longleftarrow$): Suppose that $A = \bigcup_{i \in I} B_i$ is a union of open balls B_i, where $B_i := B(x_i, r_i)$ is the open ball with center x_i and radius $r_i > 0$. We want to show that A is an open set, i.e. that each point of A is an interior point of A. So let $a \in A$ be an arbitrary point. Then there is some $j \in I$ so that $a \in B_j$, and thus $d(a, x_j) < r_j$. Choose $\varepsilon > 0$ sufficiently small so that $d(a, x_j) + \varepsilon$ is still $< r_j$. (E.g. take $\varepsilon := \frac{1}{2}(r_j - d(a, x_j)$. Then $d(a, x_j) + \varepsilon = \frac{1}{2}(d(a, x_j) + r_j) < \frac{1}{2}(r_j + r_j)$ — but draw a picture!) We now claim that $B(a, \varepsilon) \subseteq B(x_j, r_j)$: For if $z \in B(a, \varepsilon)$, then $d(z, a) < \varepsilon$, so

$$d(z, x_j) \leq d(z, a) + d(a, x_j) < \varepsilon + d(a, x_j) < r_j$$

and hence $z \in B(x_j, r_j)$. It now follows that a is indeed an interior point of A: $B(a, \varepsilon) \subseteq B(x_j, r_j) \subseteq \bigcup_{i \in I} B(x_i, r_i) = A$. Since a was an arbitrary point of A, we conclude that every point of A is an interior point of A, and thus that A is an open set.

⊣

An open ball is a "union" of a family containing just one open ball — namely itself. Hence we may conclude that open balls are open sets. The empty set is also a "union" of a family of open balls — namely no open balls — and hence the empty set is open.

Restricting to $\mathbb{R}$, it follows that set $A \subseteq \mathbb{R}$ is open if and only if A is a union of open intervals.

Proposition 3.3.4 *Let (X, d) be a metric space. A subset $A \subseteq X$ is open if and only if it contains none of its boundary points, i.e.*

$$A \text{ is open} \qquad \Longleftrightarrow \qquad A \cap \partial A = \emptyset$$

Exercise 3.3.5 Prove Proposition 3.3.4.
Hint: Suppose that $A \subseteq X$, and that $x \in A$. Show that x is an interior point of A iff it is not a boundary point of A, i.e. that

$$x \in A^\circ \qquad \Longleftrightarrow \qquad x \notin \partial A$$

□

Here are some important properties of the family of open sets in a metric space:

Proposition 3.3.6 *Let (X, d) be a metric space. The family of open sets satisfies the following axioms:*

(T.1) X and $\emptyset$ are open.

(T.2) The union of any (possibly infinite) collection of open sets is open.

(T.3) The intersection any finite collection of open sets is open.

Exercise 3.3.7 Prove Proposition 3.3.6.

□

Remarks 3.3.8 The above axioms form the basis for a further level of abstraction, i.e. a level even more general than metric spaces: A *topological space* is a set X equipped with a family $\mathcal{O}$ of subsets of X — called the open sets — satisfying T.1, T.2, T.3. The subject of *topology* studies topological spaces (and the continuous maps between them).

□

Here is a nifty topological criterion for convergence which does not mention $\varepsilon > 0$: Recall that if $\langle x_n \rangle$ is a sequence in X, then we define convergence in X by

$$x_n \to x \qquad \Longleftrightarrow \qquad d(x_n, x) \to 0$$

Proposition 3.3.9 *Let (X, d) be a metric space, and suppose that $\langle x_n \rangle$ is a sequence in X, and $x \in X$. Then $x_n \to x$ if and only if given any neighbourhood U of x, we have $x_n \in U$ eventually, i.e. for any neighbourhood U of x there is N such that $x_n \in U$ for all $n \geq N$.*

Proof: Recall that U is a neighbourhood of x if and only if there is $\varepsilon > 0$ such that $B(x, \varepsilon) \subseteq U$. Then $B(x, \varepsilon)$ is itself a neighbourhood of x.
Now suppose that $x_n \to x$, and that U is a neighbourhood of x with $B(x, \varepsilon) \subseteq U$. Then there is N such that $d(x_n, x) < \varepsilon$ for all $n \geq N$. Hence $x_n \in B(x, \varepsilon)$ for all $n \geq N$, and thus $x_n \in U$ for all $n \geq N$, i.e. $x_n \in U$ eventually.
Conversely, suppose that x_n eventually belongs to any neighbourhood of x. Let $\varepsilon > 0$. Then $x_n \in B(x, \varepsilon)$ eventually, so there is N such that $x_n \in B(x, \varepsilon)$ for any $n \geq N$. This means that for any $\varepsilon > 0$ there is N such that $d(x_n, x) < \varepsilon$ for all $n \geq N$, i.e. that $d(x_n, x) \to 0$.

⊣

3.3.3 Closed Sets

Recall that a subset A of a metric space is *defined* to be closed if and only if its complement A^c is an open set. Using de Morgan's laws and Proposition 3.3.6, it is easy to prove the analogues of (T.1)–(T.3) for closed sets:

Proposition 3.3.10 *Let X be a metric space.*

1. X and $\emptyset$ are closed.

2. The intersection of a (possibly infinite) collection of closed sets is closed.

3. The union of finitely many closed sets is closed.

Exercise 3.3.11 Prove Proposition 3.3.10 by combining Proposition 3.3.6 with De Morgan's Laws.

□

We now show that *closed* means *closed under limits*:

Proposition 3.3.12 *Suppose that (X, d) is a metric space, and that $C \subseteq X$. Then the following are equivalent:*

(i) C is closed.

(ii) Whenever $\langle c_n \rangle_n$ is a sequence in C which converges, then it converges to a point in C, i.e.

$$c_n \in C \text{ and } c_n \to x \qquad \text{implies} \qquad x \in C$$

Proof: (i) ⇒ (ii): Suppose C is closed in X, and that $\langle c_n \rangle_n$ is a sequence in C which converges, i.e. $c_n \to x$. We must show $x \in C$, and we argue by *contradiction*: If $x \notin C$, then $x \in C^c$. Since C is closed, C^c is open, so there is $\varepsilon > 0$ so that $B(x, \varepsilon) \subseteq C^c$. Since $c_n \to x$, we have $d(c_n, x) < \varepsilon$ eventually (i.e. $\exists N \in \mathbb{N}\ \forall n \geq N\ [d(c_n, x) < \varepsilon]$). Then $c_n \in B(x, \varepsilon) \subseteq C^c$ eventually, contradicting the assumption that $c_n \in C$ for all n.
(ii) ⇒ (i): We prove the *contrapositive*, i.e. we show that ¬(i) ⇒ ¬(ii): Suppose that C is not closed, i.e. that C^c is not open. Then there is a point x in C^c which is not an interior point of C^c. Thus for every $r > 0$ we have $B(x, r) \not\subseteq C^c$. Let $r_n > 0$ be real numbers such that $r_n \to 0$ (e.g. define $r_n := \frac{1}{n}$). Since $B(x, r_n) \not\subseteq C^c$, we must have $B(x, r_n) \cap C \neq \emptyset$. Choose therefore, $c_n \in B(x, r_n) \cap C$. Then $\langle c_n \rangle$ is a sequence in C, and $d(c_n, x) < r_n$, so $c_n \to x$. Yet $x \notin C$.

⊣

Proposition 3.3.13 *Let (X, d) be a metric space. A subset $C \subseteq X$ is closed if and only if it contains allCof its boundary points, i.e.*

$$A \text{ is closed} \qquad \Longleftrightarrow \qquad \partial C \subseteq C$$

Exercise 3.3.14 Prove Proposition 3.3.13.
[Hint: Note that C and C^c have the same boundary, i.e. $\partial(C^c) = \partial C$. Now apply Proposition 3.3.4 to deduce that C is closed ⇔ C^c is open ⇔ $C^c \cap \partial(C^c) = \emptyset \Leftrightarrow C^c \cap \partial C = \emptyset \Leftrightarrow \partial C \subseteq C$.]

□

Definition 3.3.15 Let (X, d) be a metric space, and $A \subseteq X$, $x_0 \in X$. We say that x_0 is a *cluster point* of A if and only if for any $r > 0$ there is $a \in A$ such that $0 < d(x_0, a) < r$, i.e. iff for every $r > 0$, there is $a \in B(x_0, r) \cap A$ such that $a \neq x_0$.
If $x_0 \in A$, but is not a cluster point of A, it is said to be an *isolated point* of A.

Note that a cluster point of a set A need not be an element of A. Note also that imposing the condition $0 < d(x_0, a) < r$ is equivalent to saying that a, x_0 are "close" (within r of each other), but not equal.

Examples 3.3.16 1. 0 is the only cluster point of the set $\{\frac{1}{n} : n \in \mathbb{N}\}$. Each element of the set is isolated.

2. x is a cluster point of the interval (a, b) if and only if $x \in [a, b]$. Thus (a, b) has no isolated points.

3. Each $x \in \mathbb{R}$ is a cluster point of the set $\mathbb{Q}$.

4. The set $\mathbb{Z}$ has no cluster points in $\mathbb{R}$.

5. A finite subset of $\mathbb{R}$ has no cluster points, i.e. each element is isolated.

6. If $x_n \to x$ and $x_n \neq x$ infinitely often, then x is a cluster point of the set $\{x_n : n \in \mathbb{N}\}$.

□

Exercise 3.3.17 (a) Find all the cluster points of the following subsets of $\mathbb{R}$ (equipped with the usual metric):

$$A := (0, 1) \qquad B := [0, 1] \qquad C := (0, 1) \cap \{2\} \qquad D := \{\tfrac{1}{n} : n \in \mathbb{N}\} \qquad E := \mathbb{Q}$$

(b) Find the boundaries $\partial A, \ldots, \partial E$ of the sets $A, \ldots, E$ above.

(c) Find the closures $\bar{A}, \ldots, \bar{E}$ of the sets $A, \ldots, E$ above.

□

Proposition 3.3.18 *Let (X, d) be a metric space, and let $A \subseteq X$. A point $x \in X$ is cluster point of A if and only if there is a sequence $\langle a_n \rangle$ with distinct terms (i.e. $n \neq m$ implies $a_n \neq a_m$) such that $a_n \to x$.*

Proof: Suppose that x is a clusterpoint of A. Then there is $a_1 \in \cap B(x, 1)$, such that $a_1 \neq x$. Let $r_1 := \min\{d(x, a_1), 1\} > 0$. Then there is $a_2 \in A \cap B(x, r_1)$ such that $a_2 \neq x$. Then $d(x, a_2) < r_1 = d(x, a_1)$, so $a_2 \neq a_1$. Next, let $r_2 := \min\{d(x, a_2), \frac{1}{2}\} > 0$. Then there is $a_3 \in A \cap B(x, r_2)$ such that $a_3 \neq x$. Then $d(x, a_2) < r_2 < r_1$, so $a_3 \neq a_2, a_1$.

Proceed inductively: Suppose we have found $a_1, \ldots, a_n$ with $d(x, a_j) = r_j$, where $1 > r_1 > r_2 > \cdots > r_{n-1} > 0$. Let $r_n := \min\{d(x, a_n), \frac{1}{n}\}$. Then there is $a_{n+1} \in A \cap B(x, r_n)$ such that $a_{n+1} \neq x$. Thus $0 < d(x, a_{n+1}) < r_n < r_{n-1} < \cdots < r_1 < 1$, so $a_{n+1} \notin \{a_1, a_2, \ldots, a_n\}$.

In this way we obtain a sequence $\langle a_n \rangle$ of distinct elements of A. Moreover, $d(x, a_n) < r_n \leq \frac{1}{n}$ for each n, and hence $d(x, a_n) \to 0$, which means $a_n \to x$.

⊣

Proposition 3.3.19 *Let X be a metric space. A subset of X is closed if and only if it contains all its cluster points.*

Exercise 3.3.20 We prove Proposition 3.3.19. Let (X, d) be a metric space. Recall that the following are equivalent for a set $C \subseteq X$.

(i) C is closed.

(ii) C contains all its boundary points.

(iii) Whenever $\langle c_n \rangle$ is a sequence in C which is convergent, then $\lim_n c_n \in C$.

(a) Suppose first that $C \subseteq X$ is closed, and that x is a cluster point of C. Explain why there is a sequence $\langle c_n \rangle$ in C such that $c_n \to x$. Conclude that C contains all its cluster points.

(b) Show that if $A \subseteq X$, and if $x \in X$ is a point such that $x \in \partial A$ but $x \notin A$, then x is a cluster point of A.

(c) Conclude that if a set contains all its cluster points, then it contains all its boundary points, and hence is closed.

□

Exercise 3.3.21 Let $X := \mathcal{C}[0,1]$ be the set of all continuous functions $[0,1] \xrightarrow{f} \mathbb{R}$.

(a) Define $d : X \times X \to \mathbb{R}$ by

$$d(f,g) := \sup\{|f(x) - g(x)| : x \in [0,1]\}$$

(i) Show that d is a metric on X. [You may assume that $\sup_{0 \le x \le 1} |f(x) - g(x)|$ is always finite when f, g are continuous — we will prove this later.]

(ii) Hence describe the open ball $B(x^2, 1)$ of radius 1 centered on the function $y = x^2$ (restricted to $[0,1]$).

(iii) Show that $f_n \to f$ in (X,d) if and only if

$$\forall \varepsilon > 0 \; \exists N \; \forall n \ge N \; \forall x \in X \; [|f_n(x) - f(x)| < \varepsilon]$$

If $f_n \to f$ in (X,d), we say that $\langle f_n \rangle$ converges *uniformly*.

(b) Define $d : X \times X \to \mathbb{R}$ by

$$d(f,g) := \left(\int_0^1 |f(x) - g(x)|^2 \, dx \right)^{\frac{1}{2}}$$

It can be shown that d is a metric. Describe the open ball $B(0,1)$ of radius 1 centered at the constant function 0.

□

3.4 Compact Sets

Compactness is one of the most important notions in analysis and topology. Yet it is very difficult to explain where the definition comes from. In some ways, compactness as a generalization of *finiteness*. Because the notion is so unfamiliar, we will restrict ourself to $\mathbb{R}$, rather than general metric spaces.

Definition 3.4.1 We say that a subset $K \subseteq \mathbb{R}$ is *sequentially compact* if and only if it has the following property: Every sequence $\langle x_n \rangle$ in K has a subsequence $\langle x_{n_k} \rangle_k$ such that

(i) $\langle x_{n_k} \rangle_k$ is convergent, and

(ii) $\lim_k x_{n_k} \in K$.

Exercise 3.4.2 (a) Show that every finite subset of $\mathbb{R}$ is sequentially compact.
[Hint: Explain why any sequence in a finite set must have a constant subsequence.]

(b) Show that $\mathbb{R}$ is not sequentially compact.

(c) Show that $(0, 1)$ is not sequentially compact.

(d) According to the Bolzano–Weierstrass Theorem, every bounded sequence in $\mathbb{R}$ has a convergent subsequence. Use this to show that every closed and bounded subset of $\mathbb{R}$ is sequentially compact.
(Note that any finite subset of $\mathbb{R}$ is closed and bounded, so (d) implies (a).)

□

Theorem 3.4.3 *Let $K \subseteq \mathbb{R}$. The following are equivalent:*

(i) K is sequentially compact: Every sequence in K has a subsequence that converges to a limit that is also in K.

(ii) K is closed and bounded.

Proof: (i) ⇒ (ii): Recall first that if a sequence converges, then so does every one of its subsequences, and to the same limit. Furthermore, a set is closed if it is closed under limits.

Suppose now that every sequence in K has a subsequence which converges to a limit in K. If $\langle k_n \rangle$ is a sequence in K with $k_n \to x$, then every subsequence of $\langle k_n \rangle$ converges to x also. Hence $x \in K$, i.e. K is closed under limits, and thus closed.

It is straightforward to show that K is also bounded: If not, we could find , for every $n \in \mathbb{N}$, a $k_n \in K$ such that $|k_n| > n$. Then no subsequence of $\langle k_n \rangle$ converges at all.

(ii) ⇒ (i): This is Exercise 3.4.2(d): Suppose that K is closed and bounded, and let $\langle k)n \rangle$ be a sequence in K. By the Bolzano–Weierstrass Theorem, any bounded sequence has a convergent subsequence, so $\langle k_n \rangle$ has a convergent subsequence $\langle k_{n_j} \rangle_j$. Since K is closed, it is closed under limits, i.e. $\lim_j k_{n_j} \in K$.

⊣

Definition 3.4.4 (a) Let $A \subseteq \mathbb{R}$. An *open cover* of A is a family $\{U_i : i \in I\}$ of open sets in $\mathbb{R}$ such that

$$A \subseteq \bigcup_{i \in I} U_i$$

(b) We say that a subset $K \subseteq \mathbb{R}$ is *compact* if and only if it has the following property: Whenever $\mathcal{U} = \{U_i : i \in I\}$ is a family of open sets such that $K \subseteq \bigcup_{i \in I} U_i$, there exists a finite subfamily $U_{i_1}, \ldots, U_{i_n} \in \mathcal{U}$ such that $K \subseteq \bigcup_{k=1}^{n} U_{i_k}$.
Succinctly put: K is compact if and only if *every* open cover of K has a finite subcover.

Examples 3.4.5 (a) Every *finite* subset of $\mathbb{R}$ is compact— why?

(b) The set $\mathbb{R}$ (with the usual metric) is *not* compact. For example, if $U_n = B(0, n)$, then $\{U_n : n \in \mathbb{N}\}$ is an open cover of $\mathbb{R}^n$. Yet it clearly has no finite subcover — why not? The same argument shows that no unbounded subset of $\mathbb{R}^n$ can be compact, i.e. compact subsets of $\mathbb{R}^n$ are necessarily bounded.

(c) No open interval (a, b) is compact in $\mathbb{R}$: Let $U_n = (a + \frac{1}{n}, b - \frac{1}{n})$. Then clearly $(a, b) = \bigcup_n U_n$ (i.e. $\{U_n\}_n$ is an open cover of (a, b)). Yet $\{U_n\}_n$ clearly has no finite subcover of (a, b) — why not?

□

Exercise 3.4.6 We prove that the closed unit interval $[0, 1]$ is a compact subset of $\mathbb{R}$.

(a) Let $I = [0, 1]$ be the closed unit interval, and let $\mathcal{U} = \{U_\gamma : \gamma \in \Gamma\}$ be an open cover of I. Define I^* to be the set of all those $x \in I$ for which $[0, x]$ can be covered by a finite subfamily of $\mathcal{U}$:

$$I^* = \{x \in [0, 1] : \exists \gamma_1, \ldots, \gamma_m \in \Gamma \ \ ([0, x] \subseteq U_{\gamma_1} \cup \cdots \cup U_{\gamma_m})\}$$

(b) Explain why $0 \in I^*$.

(c) Show that I^* is a subinterval of I: If $x \in I^*$ and $0 \leq y \leq x$, then $y \in I^*$.

(d) Define $x^* = \sup I^*$. Explain why $0 \leq x^* \leq 1$.

(e) Explain why there is $\gamma^* \in \Gamma$ such that $x^* \in U_{\gamma^*}$.

(f) Explain why $x^* \in I^*$.

(g) Assume now that $x^* < 1$. Explain why there is $\varepsilon > 0$ such that $[x^* - \varepsilon, x^* + \varepsilon] \subseteq U_{\gamma^*} \cap [0, 1]$.

(h) Explain why $[0, x^* + \varepsilon]$ can be covered by a finite subfamily of $\mathcal{U}$.

(i) Conclude that $x^* + \varepsilon \in I^*$.

(j) Explain why this is a contradiction.

(k) Deduce that $1 \in I^*$, and thus that I can be covered by a finite subfamily of $\mathcal{U}$.

□

The above exercise can easily be generalized to show that any closed interval $[a, b]$ in $\mathbb{R}$ is compact.

Theorem 3.4.7 *Let $K \subseteq \mathbb{R}$. The following are equivalent:*

(i) K is sequentially compact.

(ii) K is closed and bounded.

(iii) K is compact.

Proof: (i) $\Leftrightarrow$ (ii) is Theorem 3.4.3.

(iii) $\Rightarrow$ (ii): Suppose now that every open cover of K has a finite subcover. We first show that K is bounded. Let $r > 0$, and define $U_x = (x - r, x + r)$. Then $\{U_x : x \in K\}$ is an open cover for K (since if $y \in K$, then $y \in U_y \subseteq \bigcup_{x\in K} U_x$). By hypothesis, there is a finite subcover, i.e. there are $x_1, \ldots, x_n \in K$ such that $K \subseteq \bigcup_{i=1}^n U_{x_i}$. Hence K is contained in a union of finitely many open intervals, each of finite length, which clearly implies that K is bounded.

Next, we show that K is closed, i.e. closed under limits. So suppose that $\langle y_n\rangle$ is a convergent sequence in K, and that $y_n \to y$. We must show that $y \in K$. Now **if** $y \notin K$, define an open cover of K as follows: For each $x \in K$, let $r_x := \frac{1}{2}|x - y|$ be half the distance between x and y, and define $V_x := B(x, r_x) = (x - r_x, x + r_x)$ to be the interval of radius r_x centered at x. Then $\{V_x : x \in K\}$ is clearly an open cover of K. By assumption, there exist $x_1, \ldots, x_m \in K$ such that $K \subseteq \bigcup_{j=1}^m V_{x_j}$. Define $\varepsilon := \min\{r_{x_j} : j = 1, \ldots, m\}$. Then $\varepsilon > 0$. Note that if $x \in K$, then $|x - y| > \varepsilon$: For if $x \in K$, then there is j such that $x \in V_{x_j}$, i.e. such that $|x - x_j| < r_{x_j}$. Using the fact that $|a - b| \geq |a| - |b|$, we see that

$$|x - y| = |(x - x_j) - (y - x_j)| \geq |y - x_j| - |x - x_j| > |y - x_j| - r_{x_j} = r_{x_j} \geq \varepsilon$$

Now since $\langle y_n\rangle$ is a sequence in K, we have $|y_n - y| > \varepsilon$ for all n. Hence it is impossible that $y_n \to y$ — contradiction. Hence the **if** above leads to contradiction, and we may conclude that $y \in K$, as required.

(i) $\Rightarrow$ (iii): Suppose that $K \subseteq \mathbb{R}$ is such that every sequence in K has a subsequence that converges to a limit in K. Let $\mathcal{U} := \{U_i : i \in I\}$ be an open cover of K. Define a function $K \xrightarrow{f} \mathbb{R}$ by

$$f(x) := \sup\{r : U \in \mathcal{U} : B(x, r) \subseteq U\}$$

Since $\mathcal{U}$ covers K, each $x \in K$ belongs to some $U \in \mathcal{U}$. Since this U is open, x is an interior point of U. Hence $f(x) > 0$ for each $x \in K$. But we can say more:

$$\inf_{x\in K} f(x) > 0 \qquad (*)$$

For suppose $(*)$ is false. Then there exists a sequence $\langle y_n\rangle$ in K such that $f(y_n) \to 0$ — just pick $y_n \in K$ so that $f(y_n) < \frac{1}{n}$. By assumption, $\langle y_n\rangle$ has a convergent subsequence $\langle z_n\rangle$ such that $z := \lim_n z_n \in K$. Then $\langle f(z_n\rangle$ is a subsequence of the convergent sequence $\langle f(y_n)\rangle$ and hence $f(z_n) \to 0$ also. Since $z \in K$, there is $U \in \mathcal{U}$ such that $z \in U$. Let $r > 0$ be such that $B(z, r) \subseteq U$. Then $z_n \in B(z, \frac{r}{2})$ eventually. Then $B(z_n, \frac{r}{2}) \subseteq B(z, r) \subseteq U$, and hence $f(z_n) \geq \frac{r}{2}$ eventually. This contradicts $f(z_n) \to 0$. Hence $(*)$ holds.

With $(*)$ now proved, we can proceed: Choose c so that $0 < c < \inf_{x\in K} f(x)$ and let $x_1 \in K$ be arbitrary. If possible, choose inductively $x_{n+1} \in K$ so that $|x_{n+1} - x_j| > c$ for all $j = 1, \ldots, n$. If this is possible for all n, one would obtain a sequence $\langle x_n\rangle_n$ in K with $|x_n - x_m| > c$ for all n, m, and such a sequence *cannot* have a convergent subsequence. Hence there is n for which it is impossible to choose x_{n+1}, and hence $K \subseteq \bigcup_{j=1}^n B(x_j, c)$. By definition of c there exists a $U_{i_j} \in \mathcal{U}$ so that $B(x_j, r) \subseteq U_{i_j}$ for $j = 1, \ldots, n$. Thus $K \subseteq \bigcup_{j=1}^n U_{i_j}$ yields a finite subcover.

$\dashv$

Remarks 3.4.8 The fact that a compact set in $\mathbb{R}$ is the same as a closed and bounded set is called the *Heine–Borel* Theorem.

□

Here comes the Completeness Axiom again:

Theorem 3.4.9 *Suppose that $K_1 \supseteq K_2 \supseteq K_3 \supseteq \ldots$ is a decreasing sequence of non–empty compact subsets of $\mathbb{R}$. Then $\bigcap_{n=1}^{\infty} K_n \neq \emptyset$.*

Exercise 3.4.10 We prove Theorem 3.4.9 using the fact that a set is compact if and only if it is closed and bounded. Suppose that $K_1 \supseteq K_2 \supseteq K_3 \supseteq \ldots$ is a decreasing sequence of non–empty compact subsets of $\mathbb{R}$.

(a) Let $u_n := \sup K_n$ and $l_n := \inf K_n$. Explain why $u_n, l_n \in K_n$.

(b) Show that $l_1 \leq l_2 \leq \cdots \leq l_n \leq \cdots \leq u_n \leq \cdots \leq u_2 \leq u_1$.

(c) Deduce that $l := \lim_n l_n$ exists.

(d) Show that the tail sequence $\langle l_n \rangle_{n \geq m} = l_m, l_{m+1}, l_{m+2}, \ldots$ is a convergent sequence in K_m, and deduce that $l \in K_m$, for all m.

(e) Hence conclude that $\bigcap_m K_m \neq \emptyset$.

□

Exercise 3.4.11 (a) Give an example of a decreasing sequence of non–empty closed sets $C_n \subseteq \mathbb{R}$ so that $\bigcap_n C_n = \emptyset$.

(b) Give an example of a decreasing sequence of non–empty bounded sets $A_n \subseteq \mathbb{R}$ so that $\bigcap_n A_n = \emptyset$.

□

Exercise 3.4.12 We prove Theorem 3.4.9 again, this time using the definition of compactness. Suppose that $K_1 \supseteq K_2 \supseteq K_3 \supseteq \ldots$ is a decreasing sequence of non–empty compact sets in $\mathbb{R}$, but that $\bigcap_{n=1}^{\infty} K_n = \emptyset$. We seek to obtain a contradiction.

(a) Define $U_n := K_n^c$ for $n \in \mathbb{N}$. Explain why $\{U_n : n \in \mathbb{N}\}$ is an open cover of K_1.

(b) Note that $U_1 \subseteq U_2 \subseteq U_3 \subseteq \ldots$. Conclude that there is $N \in \mathbb{N}$ such that $K_1 \subseteq U_N$.

(c) Hence deduce that $K_1 \cap K_N = \emptyset$, and explain why this is a contradiction.

□

Chapter 4

Limits of Functions and Continuity

4.1 Limits of Functions

We have already defined the concept of limit for sequences, i.e. we know what is meant by a statement of the form $\lim_{n\to\infty} x_n = l$. Our aim in this section is to give a similar definition for

$$\lim_{x\to x_0} f(x)$$

where $f : X \to \mathbb{R}$ and $X \subseteq \mathbb{R}$ Later in this section, we shall also look at *left limits*, and *right limits*, concepts which employ the fact that $\mathbb{R}$ is also an ordered set.

Let $X \subseteq \mathbb{R}$, and let $f : X \to \mathbb{R}$ be a function. We want to investigate what we mean when we say that

$$\lim_{x\to x_0} f(x) = y_0$$

where $x, x_0 \in X$, and $y_0 \in \mathbb{R}$. Clearly the *intention* (i.e. the intuitive content) of the above statement is that as x "gets closer and closer" to x_0, $f(x)$ "gets closer and closer" to y_0. We are *not* interested in what happens to the value of f *at* the point x_0: We already know what it is — the value of the function there is simply $f(x_0)$. We *are* interested in what happens to the values of f as we get closer and closer to x_0, without actually allowing x to be equal to x_0. Indeed, it is quite possible that $x_0 \notin X$, so that $f(x_0)$ is not even defined.

As before, when we discussed limits of sequences, we run into the problem of not having a definition of "close". But we already know how to circumvent this problem. For sequences, the statement $\lim_n x_n = l$ meant that x_n "gets closer and closer" to l as n gets bigger and bigger. We conceptualized the notion of "closer and closer" as follows:

> $\lim_n x_n = l$ provided that, given any distance $\varepsilon > 0$, x_n and l lie within ε of each other whenever n is sufficiently large.

We employ the same trick here:

> $\lim_{x\to x_0} f(x) = y_0$ provided that, given any distance $\varepsilon > 0$, the points $f(x)$ and y_0 lie within ε of each other whenever the points x and x_0 are sufficiently close (but not equal) in $\mathbb{R}$.

We must therefore find, for any distance $\varepsilon > 0$, a distance $\delta > 0$ such that if x, x_0 lie within δ (but are not equal to each other), then $f(x), y_0$ lie within ε.

Thus we *define* the meaning of $\lim_{x\to x_0} f(x) = y_0$ as follows:

For any $\varepsilon > 0$ there is a $\delta > 0$ such that if $0 < |x - x_0| < \delta$, then $|f(x) - y_0| < \varepsilon$

For this definition to make sense, we must be able to find $x \in X$ which are arbitrarily close, but not equal, to x_0. This is possible precisely when x_0 is a *cluster point* of X. We recall here the definition.

Definition 4.1.1 Let $X \subseteq \mathbb{R}$, and let $x_0 \in \mathbb{R}$. We say that x_0 is a *cluster point* of X if and only if for any $\delta > 0$ there is $x \in X$ such that $0 < |x - x_0| < \delta$.
If $x_0 \in X$ is not a cluster point of X, it is said to be an *isolated point*.

Note that a cluster point of a set X need not be an element of X. Note also that imposing the condition $0 < |x - x_0| < \delta$ is equivalent to saying that x, x_0 are "close" (within δ of each other), but not equal. For revision, we recall some examples:

Examples 4.1.2 1. 0 is the only cluster point of the set $\{\frac{1}{n} : n \in \mathbb{N}\}$. Each element of the set is isolated.

2. x is a cluster point of the interval (a, b) if and only if $x \in [a, b]$. Thus (a, b) has no isolated points.

3. Each $x \in \mathbb{R}$ is a cluster point of the set $\mathbb{Q}$.

4. The set $\mathbb{Z}$ has no cluster points in $\mathbb{R}$.

5. A finite subset of $\mathbb{R}$ has no cluster points, i.e. each element is isolated.

6. If $x_n \to x$ and $x_n \neq x$ infinitely often, then x is a cluster point of the set $\{x_n : n \in \mathbb{N}\}$.

□

Definition 4.1.3 Let $f : X \to \mathbb{R}$, let $x_0 \in \mathbb{R}$ be a cluster point of X, and and let $y_0 \in \mathbb{R}$. We say that

$$\lim_{x\to x_0} f(x) = y_0 \qquad \text{or} \qquad f(x) \to y_0 \text{ as } x \to x_0$$

if and only if: For every $\varepsilon > 0$, there is a $\delta > 0$ such that

$$|f(x) - y_0| < \varepsilon \qquad \text{whenever} \qquad x \in X \text{ and } 0 < |x - x_0| < \delta$$

In logical notation:

$$\lim_{x\to x_0} f(x) = y_0 \qquad \Longleftrightarrow \qquad \forall \varepsilon > 0\ \exists \delta > 0\ \forall x \in X\ [0 < |x - x_0| < \delta\ \to\ |f(x) - y_0| < \varepsilon]$$

Remarks 4.1.4

- For the symbol $\lim_{x\to x_0}$ to make sense, it is necessary that x_0 be a cluster point of the set X. However, it need not belong to X.
- Note that δ plays the same role in the definition of $\lim_{x\to x_0} f(x)$ as does the N in the definition of $\lim_n x_n$. In particular, δ usually depends on ε: The smaller we choose ε, the smaller we will have to take δ.
  Moreover, δ will usually also depend on x_0.

- We can define the notion of limits for functions between arbitrary metric spaces: If (X, d_X) and (Y, d_Y) are metric spaces, and $f : X \to Y$, then we say that $\lim_{x \to x_0} f(x) = y_0$ if and only if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \; [0 < d_X(x, x_0) < \delta \; \rightarrow \; d_Y(f(x), y_0) < \varepsilon]$$

  i.e. $f(x)$ can be made as close to y_0 as you like (in the space (Y, d_Y)) by taking x to be sufficiently close, but not equal, to x_0 (in the space (X, d_X)).

□

Examples 4.1.5 1. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$. Then

$$\lim_{x \to 2} f(x) = 4$$

For suppose $\varepsilon > 0$. We must find a $\delta > 0$ such that

$$|x^2 - 4| < \varepsilon \qquad \text{whenever} \qquad |x - 2| < \delta$$

Choose $\delta = \min\{\frac{\varepsilon}{5}, 1\}$. Then if $|x - 2| < \delta$, then also $|x - 2| < 1$, which implies that $1 < x < 3$, and thus $|x + 2| < 5$. Now $|x - 2| < \delta$, we have

$$|x^2 - 4| = |x + 2| \cdot |x - 2| < 5|x - 2| < 5\delta \leq \varepsilon$$

as required.
This example makes it clear that δ (usually) depends on ε.

2. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$, and let $x_0 \in \mathbb{R}$. Then

$$\lim_{x \to x_0} f(x) = x_0^2$$

For suppose $\varepsilon > 0$. We must find a $\delta > 0$ such that

$$|x^2 - x_0^2| < \varepsilon \qquad \text{whenever} \qquad |x - x_0| < \delta$$

Choose $\delta = \min\{\frac{\varepsilon}{1+2|x_0|}, 1\}$. Then if $|x - x_0| < \delta$, then also $|x - x_0| < 1$, which implies that $|x + x_0| = |x - x_0 + 2x_0| \leq |x - x_0| + 2|x_0| < 1 + 2|x_0|$ Now if $0 < |x - x_0| < \delta$, we have

$$|x^2 - x_0^2| = |x + x_0| \cdot |x - x_0| < (1 + 2|x_0|)|x - x_0| < (1 + 2|x_0|)\delta \leq \varepsilon$$

as required.
This example makes it clear that δ (usually) depends on both ε and x_0.

3. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } x \geq 0 \end{cases}$$

Let us consider $\lim_{x \to 0} f(x)$. From the graph of f, it is clear that if x is close to 0 and $x > 0$, then $f(x) = 1$. On the other hand, if x is close to 0, but $x < 0$, then $f(x) = 0$. This suggests that $\lim_{x \to 0} f(x)$ does not exist.

We now *prove* that, indeed, $\lim_{x\to 0} f(x)$ does not exist. For suppose that $y_0 \in \mathbb{R}$, and that $\lim_{x\to 0} f(x) = y_0$. Let $\varepsilon = \frac{1}{2}$. Then we should be able to find a $\delta > 0$ such that $|f(x) - y_0| < \frac{1}{2}$ whenever $0 < |x| < \delta$. In particular, $|f(\frac{\delta}{2}) - y_0| < \frac{1}{2}$ and $|f(-\frac{\delta}{2}) - y_0| < \frac{1}{2}$, as $|\pm\frac{\delta}{2}| < \delta$. Thus

$$1 = |f(\tfrac{\delta}{2}) - f(-\tfrac{\delta}{2})| \leq |f(\frac{\delta}{2}) - y_0| + |y_0 - f(-\tfrac{\delta}{2})| < \tfrac{1}{2} + \tfrac{1}{2} = 1$$

i.e. $1 < 1$, a contradiction.

4. Define

$$f(x) = \begin{cases} 0 \text{ if } x \neq 0 \\ 1 \text{ if } x = 0 \end{cases}$$

Then it is easy to see that $\lim_{x\to 0} f(x) = 0$. Indeed, given $\varepsilon > 0$, let δ be any non–zero number whatsoever. If $0 < |x| < \delta$, then $x \neq 0$, so $f(x) = 0$, and so $|f(x) - 0| < \varepsilon$, as required.
Note that $\lim_{x\to 0} f(x) = 0$, whereas $f(0) = 1$.
The point is that the value of a function at a particular x_0 may be totally unrelated to the behaviour of the function as $x \to x_0$.

□

The following important proposition shows that the the definition of the limit of a function at a point x_0 can be rephrased in terms of a limits of sequences.

Proposition 4.1.6 *Suppose that $X \subseteq \mathbb{R}$, that $f : X \to \mathbb{R}$ and that $y_0 \in \mathbb{R}$. Then the following are equivalent:*

(a) $\lim_{x\to x_0} f(x) = y_0$

(b) Whenever $\langle x_n\rangle_n$ is a sequence in X such that

(i) $\lim_n x_n = x_0$ *in* $\mathbb{R}$;

(ii) $x_n \neq x_0$ *for all* $n \in \mathbb{N}$;

then we have $\lim_n f(x_n) = y_0$.

Proof: (a) ⇒ (b): Suppose that $\lim_{x\to x_0} f(x) = y_0$, and that $\langle x_n\rangle_n$ is a sequence in X satisfying (i) and (ii) above. We must show that $f(x_n) \to y_0$, i.e. we must show that

$$\text{For all } \varepsilon > 0 \text{ there is } N \text{ such that } |f(x_n) - y_0| < \varepsilon \text{ whenever } n \geq N$$

So let $\varepsilon > 0$. Because $\lim_{x\to x_0} f(x) = y_0$, there is $\delta > 0$ such that

$$0 < |x - x_0| < \delta \quad \text{implies} \quad |f(x) - y_0| < \varepsilon$$

Now because $\lim_{n\to\infty} x_n = x_0$, there is $N \in \mathbb{N}$ such that

$$n \geq N \quad \text{implies} \quad |x_n - x_0| < \delta$$

Then if $n \geq N$, we see that $0 < |x_n - x_0| < \delta$, so that $|f(x_n) - y_0| < \varepsilon$, as required. Thus (a) implies (b).

(b) $\Rightarrow$ (a): To prove the converse, we argue by contradiction: Suppose that $\lim_{x \to x_0} f(x) \neq y_0$. (In fact, the limit need not even exist.) I.e. suppose that

$$\neg \Big(\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in X \; [0 < |x - x_0| < \delta \; \rightarrow \; |f(x) - y_0| < \varepsilon] \Big)$$

which is to say, that

$$\exists \varepsilon > 0 \; \forall \delta > 0 \exists x \in X \; [|0 < x - x_0| < \delta \; \wedge \; |f(x) - y_0| \geq \varepsilon]$$

i.e. that there is a $\varepsilon > 0$ such that for every $\delta > 0$, we are able to find an $x \in X$ such that

$$0 < |x - x_0| < \delta \quad \text{but} \quad |f(x) - y_0| \geq \varepsilon$$

In particular, if $\delta = \frac{1}{n}$, we are able to find an $x_n \in X$ such that

$$0 < |x_n - x_0| < \tfrac{1}{n} \quad \text{yet} \quad |f(x) - y_0| \geq \varepsilon$$

We have thus found a sequence $\langle x_n \rangle_n$ in X satisfying (i), (ii) such that $f(x_n) \not\to y_0$. Hence not-(a) implies not-(b), i.e. (b) implies (a).

$\dashv$

Corollary 4.1.7 *(a) If $\lim_{x \to x_0} f(x)$ exists, then the limit is unique;*

(b) $\lim_{x \to x_0} f(x) + \lim_{x \to x_0} g(x) = \lim_{x \to x_0} \big(f(x) + g(x) \big)$*;*

(c) $\lim_{x \to x_0} \alpha f(x) = \alpha \lim_{x \to x_0} f(x)$*;*

(d) $\big(\lim_{x \to x_0} f(x) \big) \big(\lim_{x \to x_0} g(x) \big) = \lim_{x \to x_0} f(x)g(x)$ *when f, g are between $\mathbb{R}$ and $\mathbb{R}$;*

(e) $\frac{\lim_{x \to x_0} f(x)}{\lim_{x \to x_0} g(x)} = \lim_{x \to x_0} \frac{f(x)}{g(x)}$*, when f, g are between $\mathbb{R}$ and $\mathbb{R}$, provided that $\lim_{x \to x_0} g(x) \neq 0$.*

Proof: These properties hold for limits of sequences.

$\dashv$

.

We end this section with a brief look at *left-* and *right* limits of functions:

Definition 4.1.8 Suppose that $X \subseteq \mathbb{R}$, and that $f : X \to \mathbb{R}$. We say that $f(x)$ *tends to* y_0 *as* x *tends to* x_0 *from the right*, and write

$$\lim_{x \to x_0^+} f(x) = y_0 \qquad \text{or} \qquad f(x) \to y_0 \text{ as } x \downarrow x_0$$

if and only if: For every $\varepsilon > 0$, there is a $\delta > 0$ such that

$$|f(x) - y_0| < \varepsilon \qquad \text{whenever} \qquad 0 < x - x_0 < \delta$$

Similarly, we say that $f(x)$ *tends to* y_0 *as* x *tends to* x_0 *from the left*, and write

$$\lim_{x \to x_0^-} f(x) = y_0 \qquad \text{or} \qquad f(x) \to y_0 \text{ as } x \uparrow x_0$$

if and only if: For every $\varepsilon > 0$, there is a $\delta > 0$ such that

$$|f(x) - y_0| < \varepsilon \qquad \text{whenever} \qquad 0 < x_0 - x < \delta$$

An example will make these ideas clear:

Example 4.1.9 Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x - 1 \text{ if } x < 1 \\ 1 \text{ if } x = 1 \\ 1 + x^2 \text{ if } x > 1 \end{cases}$$

Then $\lim_{x \to 1^-} f(x) = 0$, and $\lim_{x \to 1^+} f(x) = 2$. Moreover, $\lim_{x \to 1} f(x)$ does not exist.

To see this, let $\varepsilon > 0$, and choose $\delta = \varepsilon$. Then if $1 - \delta < x < 1$, we also have $1 - \varepsilon < x < 1$. Now since $f(x) = x - 1$ when $x < 1$, we see that

$$0 < 1 - x < \delta \qquad \text{implies} |f(x) - 0| < \varepsilon$$

Thus $f(x) \to 0$ as $x \uparrow 1$.

Similarly, given $\varepsilon > 0$, choose $\delta > 0$ sufficiently small that $\delta(\delta + 2) < \varepsilon$. Note that if $0 < x - 1 < \delta$, then $|f(x) - 2| = |x^2 - 1| = (x - 1)(x + 1) < \delta(\delta + 2)$, and thus

$$0 < x - 1 < \delta \quad \Rightarrow |f(x) - 2| < \varepsilon$$

This proves that $f(x) \to 2$ as $x \downarrow 1$.

Finally, the fact that $\lim_{x \to 1} f(x)$ does not exist follows from the next proposition, Proposition 4.1.11.

□

Exercise 4.1.10 Show that if $f : \mathbb{R} \to \mathbb{R}$, then the following are equivalent:

(a) $\lim_{x \to x_0^+} f(x) = y_0$

(b) For every *strictly decreasing sequence* $x_n \to x_0$ it is the case that $f(x_n) \to y_0$.

□

The limit of a function exists if and only if both the left limit and the right limit exist, and are equal:

Proposition 4.1.11 *Suppose that $X \subseteq \mathbb{R}$, and that $f : X \to \mathbb{R}$. Then $\lim_{x \to x_0} f(x)$ exists if and only if*

(i) Both $\lim_{x \to x_0^+} f(x)$ and $\lim_{x \to x_0^-} f(x)$ exist, and

(ii) $\lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^-} f(x)$

In that case,

$$\lim_{x \to x_0} f(x) = \lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^-} f(x)$$

Exercise 4.1.12 Prove Proposition 4.1.11.

□

4.2 Continuity

In this section, we define the notion of a *continuous function*, and provide several useful characterizations of this notion. We shall get the theoretical part over with as soon as possible, and then — in the next section — we will analyze several examples, with the aim of making these abstract ideas concrete.

Intuitively, to say that f is continuous at a point $x_0 \in X$ means that as x "gets closer and closer" to x_0 in X, the function values $f(x)$ "get closer and closer" to $f(x_0)$ in Y. Thus the following definition should be be devoid of mystery:

Definition 4.2.1 (a) Let $f : X \to \mathbb{R}$ be a function, where $X \subseteq \mathbb{R}$, and let $x_0 \in X$. We say that f is *continuous at* x_0 if and only if for all $\varepsilon > 0$ there is a $\delta > 0$ such that

$$|f(x) - f(x_0)| < \varepsilon \qquad \text{whenever} \qquad x \in X \text{ and } |x - x_0| < \delta$$

i.e.

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in X \; [|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \varepsilon]$$

(b) A function $f : X \to \mathbb{R}$ is said to be *continuous* if and only if it is continuous at every point in its domain.

(c) A function is said to be *discontinuous* at x_0 if and only if it is not continuous at x_0.

Remarks 4.2.2 • It ought to be clear that continuity is a concept that applies to functions between metric spaces: If $(X, d_X) \xrightarrow{f} (Y, d_Y)$ is a function between two metric spaces, then f is said to be continuous at x_0 if and only if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in X \; [d_X(x, x_0) < \delta \; \rightarrow \; d_Y(f(x), f(x_0)) < \varepsilon]$$

i.e. $f(x)$ can be made as close to $f(x_0)$ as you like (in the space (Y, d_Y)) by taking x to be sufficiently close to x_0 (in the space (X, d_X)).

- Note that $\lim_{x\to x_0} f(x) = f(x_0)$ if and only if for all $\varepsilon > 0$ there is a $\delta > 0$ such that

$$|f(x) - f(x_0)| < \varepsilon \qquad \text{whenever} \qquad 0 < |x - x_0| < \delta$$

The difference between saying that $\lim_{x\to x_0} f(x) = f(x_0)$ and saying that f is continuous at x_0 is therefore very slight: The former requires that x_0 is a cluster point of X, whereas the latter does not.

- Note that if x_0 belongs to X, but is not a cluster point of X (i.e. if x_0 is an *isolated* point of X), then the statement "$\lim_{x\to x_0} f(x) = f(x_0)$" is meaningless, but the statement "f is continuous at x_0" is well–defined, and moreover, *always true*!!

Fact: If x_0 is an isolated point of X, then any function $X \xrightarrow{f} \mathbb{R}$ is continuous at x_0

To see this, note that if $x_0 \in X$ is not a cluster point of X, then there is a $\delta > 0$, such that if $0 < |x - x_0| < \delta$ then $x \notin X$. It follows that if $x \in X$ and $|x - x_0| < \delta$, then $x = x_0$. But then, given any $\varepsilon > 0$, we see that $|x - x_0| < \delta$ implies $x = x_0$ (because $x \in X$), so that $|f(x) - f(x_0)| = 0 < \varepsilon$.

□

To summarize these last remarks:

Proposition 4.2.3 *Let $X \subseteq \mathbb{R}$, let $x_0 \in X$, and let $X \xrightarrow{f} \mathbb{R}$.*

- *If x_0 is a cluster point of X, then f is continuous at x_0 if and only if $\lim_{x\to x_0} f(x) = f(x_0)$.*
- *If x_0 is not a cluster point of X, then f is continuous at x_0 (no matter what f is).*

Next, we discuss two characterizations of continuity: The first is in terms of sequences, and the second invokes topological notions. For this purpose, recall that

Proposition 4.2.4 *The following are equivalent:*

- *$x_n \to x_0$ in $\mathbb{R}$;*
- *Whenever a set $U \subseteq \mathbb{R}$ is a neighbourhood of x_0 (i.e. whenever x_0 is an interior point of U), then $x_n \in U$ eventually (i.e. there is N such that $x_n \in U$ for all $n \geq N$).*

Theorem 4.2.5 *Let $f : X \to \mathbb{R}$ be a function, where $X \subseteq \mathbb{R}$, and let $x_0 \in X$. Then the following are equivalent.*

(a) f is continuous at x_0.

(b) If $\langle x_n \rangle_n$ is a sequence in X such that $x_n \to x_0$, then $f(x_n) \to f(x_0)$.

(c) For every neighbourhood V of $f(x_0)$ there is a neighbourhood U of x_0 such that:

$$f[X \cap U] \subseteq V$$

Proof: (a) ⇒ (c): Suppose that $X \xrightarrow{f} \mathbb{R}$ is continuous at x_0, and that V is a neighbourhood o $f(x_0) \in V$. Then $f(x_0)$ is an interior point of V, i.e. there is $\varepsilon > 0$ such that $B(f(x_0), \varepsilon) := (f(x_0) - \varepsilon, f(x_0) + \varepsilon)$ is a subset of V. Choose now a $\delta > 0$ such that $|x - x_0| < \delta$ implies $|f(x_0) - f(x)| < \varepsilon$, and let $U := B(x_0, \delta) := (x_0 - \delta, x_0 + \delta)$. Then

$$x \in U \cap X \quad \Rightarrow \quad x \in X \wedge |x - x_0| < \delta \quad \Rightarrow \quad |f(x) - f(x_0)| < \varepsilon \quad \Rightarrow \quad f(x) \in B(f(x_0), \varepsilon) \subseteq V$$

Thus U is a neighbourhood of x_0, and $f[U \cap X] \subseteq V$.
(c) ⇒ (b): Suppose that $x_n \to x_0$ in X. We must show that $f(x_n) \to f(x_0)$, i.e. that if V is any neighbourhood of $f(x_0)$, then $f(x_n) \in V$ eventually. So let V be a neighbourhood of $f(x_0)$. By assumption, there is a neighbourhood U of x_0 such that $f(x) \in V$ whenever $x \in U \cap X$. Since $x_n \to x_0$, we must have $x_n \in U$ eventually. Hence $f(x_n) \in V$ eventually, as well.
(b) ⇒ (a): We prove this by contradiction. Suppose, therefore, that f is not continuous at x_0, i.e. that

$$\neg\Big(\forall \varepsilon > 0\ \exists \delta > 0\ \forall x \in X\ [|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \varepsilon]\Big)$$

i.e. that

$$\exists \varepsilon > 0\ \forall \delta > 0\ \exists x \in X\ [|x - x_0| < \delta \wedge |f(x) - f(x_0)| \geq \varepsilon]$$

Thus there is an $\varepsilon > 0$ such that for every $\delta > 0$ we can find an $x \in X$ with the properties that $|x - x_0| < \delta$, yet $|f(x) - f(x_0)| \geq \varepsilon$. Fix such an $\varepsilon > 0$, and successively take $\delta = \frac{1}{n}$ for $n \in \mathbb{N}$. This yields, for each n, an $x_n \in X$ such that $|x_n - x_0| < \frac{1}{n}$, yet $|f(x_n) - f(x_0)| \geq \varepsilon$. Then $x_n \to x_0$ in X, yet $f(x_n) \not\to f(x_0)$ in $\mathbb{R}$. Thus not-(a) implies not-(b).

⊣

Up to now we have discussed continuity *at a point*. Here is a beautiful characterization of *global continuity*:

Theorem 4.2.6 *Let $f : \mathbb{R} \to \mathbb{R}$. Then the following are equivalent.*

(1.) f is (everywhere) continuous.

(2.) Pullbacks of open sets are open: Whenever $U \subseteq \mathbb{R}$ is an open set, then its inverse image $f^{-1}[U]$ is open also.

(3.) Pullbacks of closed sets are closed: Whenever $C \subseteq \mathbb{R}$ is an closed set, then its inverse image $f^{-1}[C]$ is closed also.

In abstract topology, (b) is used as a *definition* of continuous function! It uses only the notion of open sets: No ε, no metric.

Exercise 4.2.7 We prove Theorem 4.2.6.

(a) We first prove that (1.) ⇒ (2.): Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous, and that $U \subseteq \mathbb{R}$ is open. We want to prove that $f^{-1}[U]$ is open, i.e. that every point of $f^{-1}[U]$ is an interior point.

(i) So let $x_0 \in f^{-1}[U]$ Explain why there is $\varepsilon > 0$ such that $B(f(x_0), \varepsilon) \subseteq U$.

(ii) Next, explain why there is $\delta > 0$ such that $x \in B(x_0, \delta)$ implies $f(x) \in B(f(x_0), \varepsilon)$.

(iii) Conclude that $B(x_0, \delta) \subseteq f^{-1}[U]$, and thus that x_0 is an interior point of U.

(b) Next, we show (2.) $\Rightarrow$ (1.): Suppose that $f : \mathbb{R} \to \mathbb{R}$ is such that pullbacks of open sets are open. We want to prove that f is continuous at every $x_0 \in \mathbb{R}$. So let $\varepsilon > 0$.

(i) Explain why x_0 is an interior point of $f^{-1}[B(f(x_0), \varepsilon)]$.

(ii) Conclude that there is $\delta > 0$ such that $B(x_0, \delta) \subseteq f^{-1}[B(f(x_0), \varepsilon)]$.

(iii) Hence show that if $|x - x_0| < \delta$, then $|f(x_0) - f(x)| < \varepsilon$.

(iv) Conclude that f is continuous at every x_0.

(c) Use the fact that pullbacks commute with set operations to prove that (2.) $\Leftrightarrow$ (3.).

□

Finally, some definitions of one–sided continuity:

Definition 4.2.8 Let $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$, and $x_0 \in X$. We say that f is *right–continuous* at x_0 if and only if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; [0 \leq x - x_0 < \delta \to |f(x) - f(x_0)| < \varepsilon]$$

Similarly, f is said to be left–continuous at x_0 if and only if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; [0 \leq x_0 - x < \delta \to |f(x) - f(x_0)| < \varepsilon]$$

The following assertions are easy to see:

Remarks 4.2.9 Let $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$, and $x_0 \in X$.

- f is continuous at x_0 if and only if it is both left– and right–continuous at x_0.
- If x_0 is a cluster point of X, then f is right–continuous at x_0 if and only if $\lim_{x \to x_0^+} f(x) = f(x_0)$.
- Similarly, f is left–continuous at the cluster point x_0 if and only if $\lim_{x \to x_0^-} f(x) = f(x_0)$.

□

4.3 Operations on Continuous Functions; Examples

This short section lists some simple results about the preservation of continuity. The first result shows that compositions of continuous functions are continuous.

Theorem 4.3.1 *Suppose that $f : X \to Y$ and that $g : Y \to \mathbb{R}$, where $X, Y \subseteq \mathbb{R}$. Suppose further that f is continuous at x_0, and that g is continuous at $f(x_0)$. Then $g \circ f$ is continuous at x_0.*

Proof: We give three proofs, each involving a different characterization of continuity:
First Proof: Let $y_0 = f(x_0)$, and suppose that $\varepsilon > 0$. Choose $\delta_1 > 0$ such that

$$|y - y_0| < \delta_1 \quad \text{implies} \quad |g(y) - g(y_0)| < \varepsilon$$

Then (using δ_1 as your ε), choose a $\delta > 0$

$$|x - x_0| < \delta \quad \text{implies} \quad |f(x) - f(x_0)| < \delta_1$$

We then note that

$$|x - x_0| < \delta \Longrightarrow |f(x) - f(x_0)| < \delta_1 \Longrightarrow |g(f(x)) - g(f(x_0))| < \varepsilon$$

Second proof: It suffices to show that if $x_n \to x_0$, then $(g \circ f)(x_n)) \to (g \circ f)(x_0))$. Now if $x_n \to x_0$, then $f(x_n) \to f(x_0)$ because f is continuous at x_0. It follows immediately that $g(f(x_n)) \to g(f(x_0))$, because g is continuous at $f(x_0)$.
Third Proof: We show that if V is a neighbourhood of $g(f(x_0))$, then there is a neighbourhood U of x_0 such that

$$x \in X \cap U \qquad \Longrightarrow \qquad (g \circ f)(x) \in V$$

So let V be a neighbourhood of $g(f(x_0))$. Since g is continuous at $f(x_0)$, there is a neighbourhood W of $f(x_0)$ such that

$$y \in Y \cap W \qquad \Longrightarrow \qquad g(y) \in V$$

Then, because f is continuous at x_0, there is a neighbourhood U of x_0 such that

$$x \in X \cap U \qquad \Longrightarrow \qquad f(x) \in W$$

Of course, since $f : X \to Y$, each $f(x) \in Y$. Hence

$$x \in X \cap U \qquad \Longrightarrow \qquad f(x) \in Y \cap W \qquad g(f(x)) \in V$$

as required.

⊣

Suppose that $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$. To begin with, note that to say

$$\lim_{x \to x_0} f(x) = y_0$$

is exactly the same as to say

$$\lim_{h \to 0} f(x_0 + h) = y_0 \qquad (x_0 + h \in X)$$

To see this equivalence, just define $h = x - x_0$. Then $h \to 0$ if and only if $x \to x_0$, and $f(x_0 + h) = f(x)$.

Proposition 4.3.2 *The absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ is continuous.*

Proof: We use the sequence characterization of continuity. It suffices to show that if $x_n \to x$ in $\mathbb{R}$, then $|x_n| \to |x_0|$ in $\mathbb{R}$. But this we already know. (But in case you have forgotten, just note that $|\,|x_n| - |x_0|\,| \leq |x_n - x_0|$.)

$\dashv$

.

Note that if $f : X \to \mathbb{R}$, then the function $|f| : X \to \mathbb{R} : x \mapsto |f(x)|$ is simply the composition of f with $|\cdot|$, i.e. $|f| = |\cdot| \circ f$. Thus:

Corollary 4.3.3 *If $f : X \to \mathbb{R}$ is continuous, then so is $|f| : X \to \mathbb{R}$.*

The following result is a trivial consequence of the properties of limits:

Theorem 4.3.4 *If f, g are continuous at x_0, and if $\alpha \in \mathbb{R}$, then*

$$f + g, fg, \alpha f, \frac{f}{g}$$

are also continuous (when they are defined, and provided no division by zero occurs).

Of course, for $\frac{f}{g}$ to be well–defined, g must be a non–zero real–valued function.

Proposition 4.3.5 *If $p(x)$ is a polynomial with real coefficients, then $p : \mathbb{R} \to \mathbb{R}$ is continuous.*

Exercise 4.3.6 Prove Propn. 4.3.5

$\square$

Exercise 4.3.7 We prove that the function $\sqrt{\ } : \mathbb{R}^+ \to \mathbb{R}$ is continuous.

(a) Suppose that $\langle y_n \rangle_n$ is a *non–negative* sequence in $\mathbb{R}$ which converges to y. Explain why it suffices to show that then also $\sqrt{y_n} \to \sqrt{y}$.

(b) We consider, separately, two cases: $y > 0$ and $y = 0$. First assume that $y = 0$, i.e. that $y_n \to 0$. Prove that in that case also $\sqrt{y_n} \to 0$.

(c) Next assume that $y > 0$. Explain why there is a $K > 0$ such that $y \geq K$ and such that $y_n \geq K$ *eventually.*

(d) Now notice that $(\sqrt{y_n} - \sqrt{y})(\sqrt{y_n} + \sqrt{y}) = (y_n - y)$, and that $(\sqrt{y_n} + \sqrt{y}) \geq 2\sqrt{K}$. Use these facts to show that $\sqrt{y_n} \to \sqrt{y}$ in this case also.
[Hint: Choose N such that $|y_n - y| < 2\sqrt{K}\varepsilon$ whenever $n > N$.]

$\square$

Example 4.3.8 Trigonometric functions are continuous on their domain.

- To begin with, let us prove that the function $\sin(x)$ is continuous at 0. Recall the following inequality:
$$|\sin(x)| \leq |x|$$
This is easy to see geometrically: Consider a unit circle centered at the origin, and draw a ray at an angle of x radians with the positive x-axis. The ray intersects the circle at the point $(\cos(x), \sin(x))$. The length of the arc from the point $(1, 0)$ on the x–axis to the point of intersection is $|x|$. the perpendicular height of this point is $|\sin(x)|$.

- Thus a sandwich argument shows that $\lim_{x\to 0} \sin(x) = 0$, i.e. that $\sin(x)$ is continuous at $x = 0$.

- Next, note that
$$\sin(x) - \sin(x_0) = 2\sin\left(\frac{x - x_0}{2}\right)\cos\left(\frac{x + x_0}{2}\right)$$
and thus that
$$|\sin(x) - \sin(x_0)| \leq 2\left|\sin\left(\frac{x - x_0}{2}\right)\right|$$

- Let $\varepsilon > 0$, and choose $\delta_1 > 0$ such that $|h| < \delta_1$ implies $|\sin(h)| < \frac{\varepsilon}{2}$. Then choose $\delta = 2\delta_1$. It is now easy to see that
$$|x - x_0| < \delta \Longrightarrow \left|\frac{x - x_0}{2}\right| < \delta_1 \Longrightarrow \left|\sin\left(\frac{x - x_0}{2}\right)\right| < \frac{\varepsilon}{2} \Longrightarrow |\sin(x) - \sin(x_0)| < \varepsilon$$

- Now the continuity of $\cos(x)$ follows from the relationship $\sin^2(x) + \cos^2(x) = 1$ and the continuity of $\sqrt{x}$.

- All other trigonometric functions are ratios involving $\sin(x)$ and $\cos(x)$, and are thus continuous, by Theorem 4.3.4.

□

Example 4.3.9 Consider the function $f : \mathbb{Q} \to \mathbb{R}$ defined by
$$f(x) := \begin{cases} 0 & \text{if } x^2 > 2 \\ 1 & \text{if } x^2 \leq 2 \end{cases}$$
Clearly f has a "jump" at $\pm\sqrt{2}$. However, we claim that f is nevertheless continuous: after all, $\pm\sqrt{2} \notin \mathbb{Q}$. To prove it, let $x_0 \in \mathbb{Q}$, and let $\varepsilon > 0$. We must find $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $|x - x_0| < \delta$. There are three possibilities:
Case 1: $x_0 < -\sqrt{2}$. Choose $\delta > 0$ such that $|x - x_0| < \delta$ implies $x < -\sqrt{2}$. Then $|f(x) - f(x_0)| = 0 < \varepsilon$.
Case 2: $-\sqrt{2} < x_0 < \sqrt{2}$. Choose $\delta > 0$ such that $|x - x_0| < \delta$ implies $-\sqrt{2} < x < \sqrt{2}$. Then $|f(x) - f(x_0)| = 0 < \varepsilon$.
Case 3: $x_0 > \sqrt{2}$. Choose $\delta > 0$ such that $|x - x_0| < \delta$ implies $x > \sqrt{2}$. Then $|f(x) - f(x_0)| = 0 < \varepsilon$.

□

Exercise 4.3.10 Let $X = (-1, 0) \cup (0, 1)$. Define $f : X \to \mathbb{R}$ by

$$f(x) := \begin{cases} 0 & \text{if } x \in X \text{ and } x < 0 \\ 1 & \text{if } x \in X \text{ and } x > 0 \end{cases}$$

Show that f is continuous on X.

□

Example 4.3.11 We exhibit *Dirichlet's function*, a function which is discontinuous at every point.
Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational} \end{cases}$$

To see that f is everywhere discontinuous is rather easy: Fix $x_0 \in \mathbb{R}$. Choose two sequences $x_n \to x_0$ and $y_n \to x_0$ such that $\langle x_n \rangle_n$ consists only of rationals, and $\langle y_n \rangle_n$ consists only of irrationals. Then we cannot have both $\lim_n f(x_n) = f(x_0)$ *and* $\lim_n f(y_n) = f(x_0)$, because then $0 = 1$.

□

Exercise 4.3.12 We exhibit a function which is continuous at every irrational number, but discontinuous at every rational number.
Let $X = (0, \infty)$, and define $f : X \to \mathbb{R}$ as follows: If x is irrational, define $f(x) = 0$. If x is rational, write $x = \frac{m}{n}$, where m, n are relatively prime positive integers, and define $f(x) = \frac{1}{n}$.

(a) First show that f is *discontinuous* at every rational number $x > 0$.
[Hint: Consider a sequence $\langle x_n \rangle_n$ of irrational numbers converging to x.]

(b) Explain why the interval $(x - \frac{1}{2}, x + \frac{1}{2})$ contains only finitely many rational numbers which have denominators that are $\leq n$.
[Hint: The interval has length 1, so it can contain no more than 1 integer, no more than two integer multiples of $\frac{1}{2}$, no more than three integer multiples of $\frac{1}{3}$, etc.]

(c) Next we show that f is continuous at every irrational number $x > 0$. So let x be an irrational number, let $\varepsilon > 0$, and choose $n \in \mathbb{N}$ such that $\frac{1}{n} < \varepsilon$. Explain why we are able to choose a $\delta < \frac{1}{2}$ such that the interval $(x - \delta, x + \delta)$ contains no rational number with denominator $\leq n$.

(d) Conclude that $0 \leq f(y) \leq \frac{1}{n} < \varepsilon$ for every $y \in (x - \delta, x + \delta)$.

(e) Deduce that f is continuous at x if x is irrational.

□

4.4 Continuous Functions on Compact Sets

Recall that a set $K \subseteq \mathbb{R}$ is compact iff it is sequentially compact iff it is closed and bounded.

The following result is often regarded as "intuitively clear":

Theorem 4.4.1 (Intermediate Value Theorem)
Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function, and that $f(a) < f(b)$. Suppose that $y \in \mathbb{R}$ is such that $f(a) < y < f(b)$. Then there exists an $x \in [a, b]$ such that $f(x) = y$. A similar result holds if $f(b) > f(a)$.

Before we prove it, we make some remarks:

Remarks 4.4.2 Consider the function $f : \mathbb{Q} \cap [-2, 1] \to \mathbb{R}$

$$f(x) := \begin{cases} 0 & \text{if } x^2 > 2 \\ 1 & \text{if } x^2 \leq 2 \end{cases}$$

A similar function was studied in Example4.3.9, from which it is clear that f is continuous on $\mathbb{Q} \cap [-2, 1]$. However, the Intermediate Value Theorem clearly fails to hold, as $f(-2) = 0, f(1) = 1$, yet there is no $x \in \mathbb{Q} \cap [-2, 1]$ such that $f(x) = \frac{1}{2}$.
Now as Körner points out in his book *A Companion to Analysis*,

> "If this theorem is "intuitively clear" over $\mathbb{R}$, it ought to be intuitively clear over $\mathbb{Q}$.

What makes the Intermediate Value Theorem true in $\mathbb{R}$ is, of course, the Completeness Axiom. Indeed, the Completeness Axiom is, in a sense that can be made precise, equivalent to the Intermediate Value Theorem.

□

Proof: Inductively, we define two sequences $\langle a_n \rangle, \langle b_n \rangle$ such that

(i) We have

$$a_1 \leq a_2 \leq a_3 \leq \cdots \leq a_n \leq a_{n+1} \leq \cdots \leq b_{n+1} \leq b_n \leq \cdots \leq b_3 \leq b_2 \leq b_1$$

(ii) Furthermore, $f(a_n) \leq y \leq f(b_n)$ for all $n \in \mathbb{N}$

(iii) Finally, $b_n - a_n = (b - a)2^{-n+1}$ for all $n \in \mathbb{N}$.

Define $a_1 := a, b_1 := b$. Then (i), (ii), (iii) (up to $n = 1$) are clearly true.

Now suppose that a_n, b_n have been defined such that

(i)$_n$ $a_1 \leq a_2 \leq \cdots \leq a_n \leq b_n \leq \cdots \leq b_2 \leq b_1$;

(ii)$_n$ $f(a_n) \leq y \leq f(b_n)$;

(iii)$_n$ $b_n - a_n \leq (b - a)2^{-n+1}$.

Define $c_n := \frac{a_n+b_n}{2}$. There are now two possibilities:
Case I: If $f(c_n) \leq y$, define $a_{n+1} := c_n$ and $b_{n+1} := b_n$.
Case II: If $f(c_n) > y$, define $a_{n+1} := a_n$ and $b_{n+1} = c_n$.
In either case we see that

$$a_n \leq a_{n+1} \leq b_{n+1} \leq b_n$$

so that $(\text{i})_{n+1}$ holds. Moreover, certainly $f(a_{n+1}) \leq y \leq f(b_{n+1})$, so that $(\text{ii})_{n+1}$ is true. Finally,

$$(b_{n+1} - a_{n+1}) = \tfrac{1}{2}(b_n - a_n) = \tfrac{1}{2}(b-a)2^{-n+1} = (b-a)2^{-n}$$

proving $(\text{iii})_{n+1}$.

This completes the indcutive definition of the sequences $\langle a_n \rangle, \langle b_n \rangle$. Note that each $a_n, b_n \in [a,b]$. Further observe that $\lim_n (b_n - a_n) = (b-a)\lim_n 2^{-n+1} = 0$.

As $\langle a_n \rangle$ is a bounded increasing sequence (bounded above by $b = b_1$, for example), it is convergent — here is where the Completeness Axiom makes its appearance. Therefore let $x := \lim_n a_n$. Note also that $b_n = a_n + (b_n - a_n)$, so that

$$\lim_n b_n = \lim_n a_n + \lim_n (b_n - a_n) = x + 0 = x$$

i.e. $\lim_n a_n = x = \lim_n b_n$. Clearly $x \in [a,b]$ also.

Now we use the sequence criterion for continuity: Since f is continuous and $a_n \to x$, we have $f(a_n) \to f(x)$. We conclude that $f(x) \leq y$, as each $f(a_n) \leq y$. Similarly $f(b_n) \to f(x)$, from which we obtain $f(x) \geq y$. We conclude that $f(x) = y$, as required.

⊣

The following corollary is of *fundamental* importance for optimization. It, too, is often regarded as "obvious".

Theorem 4.4.3 *Suppose that $f : [a,b] \to \mathbb{R}$ is continuous, Then f* attains *its infimum and supremum on $[a,b]$, i.e. there exist $x^*, x_* \in [a,b]$ such that*

$$f(x^*) = \sup_{x \in [a,b]} f(x) \qquad f(x_*) = \inf_{x \in [a,b]} f(x)$$

Exercise 4.4.4 We prove Theorem 4.4.3.

(a) Let $K := [a,b]$. We show that f attains its supremum on K, i.e. that there is $x^* \in K$ such that $f(x^*) = \sup f[K]$. We begin by showing that

$$f[K] := \{f(x) : x \in K\}$$

is a subset of $\mathbb{R}$ which is bounded above.

a.(i) Suppose that $f[K]$ is *not* bounded above. Explain why there is a sequence $\langle x_n \rangle$ in K such that $f(x_n) \geq n$.

a.(ii) Explain why the sequence $\langle x_n \rangle$ has a convergent subsequence.

a.(iii) Explain why the sequence $\langle f(x_n) \rangle$ has a convergent subsequence.

a.(iv) Explain why this leads to contradiction.

We thus see that the set $f[K]$ must be bounded above.

(b) Explain why $y^* := \sup f[K]$ exists.

(c) Explain why there is a sequence $\langle x_n \rangle$ in K such that $f(x_n) \to y^*$.

(d) Explain why there is a subsequence $\langle x_{n_k} \rangle$ of $\langle x_n \rangle$ which converges.

(e) Define $x^* := \lim_k x_{n_k}$ to be the limit of this subsequence. Explain why $x^* \in [a, b]$.

(f) Also explain why $f(x^*) = y^*$.

(g) Where is the Completeness Axiom used in this proof?

□

In fact, we can do better than Theorem 4.4.3. Recall that if f is continuous, then the inverse image of an open set is open. Next, we show that if f is continuous, then the direct image of compact sets is compact.

Theorem 4.4.5 *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous, and that $K \subseteq \mathbb{R}$ is compact. Then $f[K]$ is compact*

Proof: Let $\mathcal{U} := \{U_i : i \in I\}$ be an open cover for $f[K]$. We must show that $\mathcal{U}$ has a finite subcover, i.e. that there are $i_1, \ldots, i_n \in I$ such that $f[K] \subseteq \bigcup_{j=1}^n U_{i_j}$. Define $V_i := f^{-1}[U_i]$ for $i \in I$. Since f is continuous, pullbacks along f preserve open sets, i.e. each V_i is open. Now note that

$$x \in K \implies f(x) \in f[K] \subseteq \bigcup_{i \in I} U_i \implies x \in f^{-1}[\bigcup_{i \in I} U_i] = \bigcup_{i \in I} V_i$$

It follows that $\mathcal{V} := \{V_i : i \in I\}$ is an open cover for K. Since K is compact, there are $i_1, \ldots, i_n \in I$ such that $K \subseteq \bigcup_{j=1}^n V_{i_j}$. Then $K \subseteq f^{-1}[\bigcup_{j=1}^n U_{i_j}]$, and thus $f[K] \subseteq \bigcup_{j=1}^n U_{i_j}$, i.e. $U_{i_1}, \ldots, U_{i_n}$ form a finite subcover of K.

⊣

We immediately obtain the following improvement of Theorem 4.4.3.

Corollary 4.4.6 *Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous,and that $K \subseteq \mathbb{R}$ is compact and non–empty. Then f* attains *its infimum and supremum on K, i.e. there exist $x^*, x_* \in K$ such that*

$$f(x^*) = \sup_{x \in K} f(x) \qquad f(x_*) = \inf_{x \in K} f(x)$$

Proof: $f[K]$ is compact, hence closed and bounded. Since $f[K]$ is bounded, $u := \sup_{x \in K} f(x)$ is finite. Since $f[K]$ is closed, $u \in f[K]$ (To see this, note that $u - \frac{1}{n}$ is not an upper bound of $f[K]$, being smaller than the smallest upper bound. Hence there is $y_n \in f[K]$ such that $u - \frac{1}{n} < y_n \leq u$, and thus y_n is a sequence in $f[K]$ with limit u. Since $f[K]$ is closed, it is closed under limits, and so $u \in f[K]$.) Thus there is $x^* \in K$ such that $f(x^*) = u$.

⊣

Chapter 5

Differentiable Functions

In this chapter we shall mainly concern ourselves with the differentiation of real–valued functions of one variable. Moreover, we shall not even begin to attempt to cover the colossal body of applications of differential calculus — that has already been accomplished in standard courses on uni– and multivariate calculus. Our aim is more modest: To provide a rigorous development of the theory behind the differential calculus.

5.1 Differentiation in $\mathbb{R}$

We shall define the notion of derivative for functions $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$. Usually, we shall assume that X is an open or closed interval, but for the moment, we would like an as general definition as possible.

Definition 5.1.1 Let $f : X \to \mathbb{R}$ be a function, where $X \subseteq \mathbb{R}$, and suppose that x_0 both belongs to X *and* is a cluster point of X. We shall say that f is *differentiable* at x_0 if and only if the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \qquad (x \in X)$$

exists. In that case, we call this limit the *derivative* of f at x_0, and denote it by $f'(x_0)$.
f is said to be differentiable if and only if it is differentiable at every point in its domain.

Remarks 5.1.2 (1.) Thus f is differentiable at x_0 if and only there exists a number L such that

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in X \; \left[0 < |x - x_0| < \delta \longrightarrow \left| \frac{f(x) - f(x_0)}{x - x_0} - L \right| < \varepsilon \right]$$

Then $f'(x_0) = L$.

(2.) x_0 is required to be a cluster point of X so that we can find $x \in X$ that are arbitrarily close to x_0 (without having $x = x_0$). x_0 is required to be an element of X so that $f(x_0)$ (which occurs in the definition of $f'(x_0)$) exists.

(3.) Note that if X is an interval, then any point that belongs to X is also a cluster point of X. Generally, therefore, we do not need to worry about the "cluster point condition".

(4.) Equivalently, if the domain X of f is an open interval, then

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

(5.) Just like we can use the order relation on $\mathbb{R}$ to define one–sided limits, we can also define one–sided derivatives. For example, if $\lim\limits_{x \to x_0^+} \frac{f(x)-f(x_0)}{x-x_0}$ exists, we call it the right–hand derivative of f at x_0.
Indeed, if $X = [a, b]$, then $f'(a)$ (as we have defined it) is just $\lim\limits_{x \to a^+} \frac{f(x)-f(a)}{x-a}$.

(6.) If f is differentiable (i.e. differentiable at every point in its domain), then f' is itself a function, $f' : X \to \mathbb{R}$. It then makes sense to ask whether f' is differentiable at a point x_0, i.e. whether $\lim\limits_{x \to x_0} \frac{f'(x)-f'(x_0)}{x-x_0}$ exists. If it does, we call this limit the *second derivative* of f at x_0, and denote it by $f''(x_0)$.
This easily generalizes to higher order derivatives. The n^{th} derivative of f at x_0 is denoted by $f^{(n)}(x_0)$.

□

Exercise 5.1.3 Show that if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^n$, where $n \in \mathbb{N}$, then $f'(x) = nx^{n-1}$.
[Hint: Put $h = x - x_0$. Then

$$x^n = x_0^n + \binom{n}{1} x_0^{n-1} h + \binom{n}{2} x_0^{n-2} h^2 + \cdots + \binom{n}{n} h^n \qquad]$$

□

Theorem 5.1.4 *Suppose that $f : X \to \mathbb{R}$ is differentiable at x_0. Then f is continuous at x_0.*

Proof: This follows easily from the properties of limits:

$$\begin{aligned} \lim_{x \to x_0} [f(x) - f(x_0)] &= \lim_{x \to x_0} \left[\frac{f(x) - f(x_0)}{x - x_0} \cdot (x - x_0) \right] \\ &= \lim_{x \to x_0} \left[\frac{f(x) - f(x_0)}{x - x_0} \right] \cdot \lim_{x \to x_0} (x - x_0) \\ &= f'(x_0) \cdot 0 = 0 \end{aligned}$$

⊣

It follows immediately that if a function is discontinuous at a point, then it cannot be differentiable there. Note, however, that the converse of Theorem 5.1.4 is false: if a function is continuous at a point, it need certainly not be differentiable there. The function $f(x) = |x|$ is easily shown to be continuous at $x_0 = 0$, but not differentiable there.

Remarks 5.1.5 Indeed, there are functions which are continuous at every point, but differentiable *nowhere*. Just try to imagine such a function!
Functions which are everywhere continuous but nowhere differentiable used to be regarded as "pathological" curiosities. These days they are used every day by physicists and actuaries to model random phenomena (such as Brownian motion, or stock prices).

□

The following representation is often useful:

Lemma 5.1.6 *Let $f : X \to \mathbb{R}$. Then f is differentiable at x_0 in X if and only if there is exists a number $a \in \mathbb{R}$ and a function $u : X \to \mathbb{R}$ such that*

(i) $u(x) \to 0$ *as* $x \to x_0$, *and*

(ii) $f(x) = f(x_0) + (x - x_0)[a + u(x)]$

In that case $a = f'(x_0)$.

Proof: Suppose that f is differentiable at x_0. Define

$$u(x) = \begin{cases} \dfrac{f(x) - f(x_0)}{x - x_0} - f'(x_0) & \text{if } x \neq x_0 \\ 0 & \text{else} \end{cases}$$

Then clearly $u(x) \to 0$ as $x \to x_0$.
Conversely, suppose there are a number $a \in \mathbb{R}$ and a function u with the properties stated above. Then $u(x) = \frac{f(x)-f(x_0)}{x-x_0} - a$ whenever $x \neq x_0$. The fact that $\lim\limits_{x \to x_0} u(x) = 0$ now easily implies that $f'(x_0)$ exists and that $f'(x_0) = a$.

⊣

We must now settle several debts that remain unpaid from a first year course in calculus:

Theorem 5.1.7 *Suppose that $f, g : X \to \mathbb{R}$ are differentiable at $x_0 \in X$. Then the functions $f + g, fg$ and $\frac{f}{g}$ are also differentiable at x_0 (assuming $g(x_0) \neq 0$ in the case of $\frac{f}{g}$), and*

(a) $(f + g)'(x_0) = f'(x_0) + g'(x_0)$

(b) $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$

(c) $\left(\dfrac{f}{g}\right)'(x_0) = \dfrac{g(x_0)f'(x_0) - g'(x_0)f(x_0)}{g(x_0)^2}$

Proof: (a) follows immediately from the properties of limits of functions.
To prove (b), define $h(x) = f(x)g(x)$. Then

$$h(x) - h(x_0) = f(x)[g(x) - g(x_0)] + g(x_0)[f(x) - f(x_0)]$$

and thus

$$\frac{h(x) - h(x_0)}{x - x_0} = f(x)\frac{g(x) - g(x_0)}{x - x_0} + g(x_0)\frac{f(x) - f(x_0)}{x - x_0}$$

Now let $x \to x_0$. Note that by Theorem 5.1.4 we have $\lim\limits_{x \to x_0} f(x) = f(x_0)$. Now use the properties of limits of functions to obtain

$$h'(x_0) = f(x_0)g'(x_0) + g(x_0)f'(x_0)$$

Finally, to prove (c), define $h(x) = \frac{f(x)}{g(x)}$, and note that

$$\frac{h(x) - h(x_0)}{x - x_0} = \frac{1}{g(x)g(x_0)} \left[g(x_0) \frac{f(x) - f(x_0)}{x - x_0} - f(x_0) \frac{g(x) - g(x_0)}{x - x_0} \right]$$

Letting $x \to x_0$ yields the result.

⊣

Theorem 5.1.8 (Chain Rule)
Suppose that

(i) $f : X \to \mathbb{R}$ *and* $g : Y \to \mathbb{R}$ *are functions, where* X, Y *are intervals in* $\mathbb{R}$ *such that* $f[X] \subseteq Y$.

(ii) f *is differentiable at* $x_0 \in X$.

(iii) g *is differentiable at* $f(x_0) \in Y$.

Then $g \circ f$ *is differentiable at* x_0, *and*

$$(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$$

Proof: Let $h = g \circ f$, and let $y_0 = f(x_0)$. Since f, g are differentiable at x_0 and y_0 respectively, there exist, by Lemma 5.1.6, two functions $u : X \to \mathbb{R}$, $v : Y \to \mathbb{R}$ such that

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)[f'(x_0) + u(x)] \\ g(y) &= g(y_0) + (y - y_0)[g'(y_0) + v(y)] \end{aligned}$$

and where $\lim_{x \to x_0} u(x) = 0$, $\lim\limits_{y \to y_0} v(y) = 0$.
Now let $y = f(x)$. It follows that

$$\begin{aligned} h(x) - h(x_0) &= g(y) - g(y_0) \\ &= (y - y_0) \cdot [g'(y_0) + v(y)] \\ &= [f(x) - f(x_0)] \cdot [g'(f(x_0)) + v(f(x))] \\ &= (x - x_0) \cdot [f'(x_0) + u(x))] \cdot [g'(f(x_0)) + v(f(x))] \\ &= (x - x_0) \cdot [g'(f(x_0))f'(x_0) + k(x)] \end{aligned}$$

where

$$k(x) = u(x)g'(f(x_0)) + v(f(x))f'(x_0) + u(x)v(f(x))$$

Now as $x \to x_0$, we see that $f(x) \to f(x_0)$ (because f is differentiable at x_0, and thus continuous there). This implies that $\lim\limits_{x \to x_0} v(f(x)) = 0$ (because $\lim\limits_{y \to y_0} v(y) = 0$, and $y_0 = f(x_0)$, with $y = f(x)$). Hence $\lim\limits_{x \to x_0} k(x) = 0$. By Lemma 5.1.6, it follows that $g \circ f$ is differentiable at x_0, and that $(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$.

⊣

Example 5.1.9 Consider the function

$$f(x) = \sin \tfrac{1}{x}$$

which is defined for all $x \neq 0$. This function has a peculiar property: It bounces between -1 and 1 infinitely often on any interval of the form $(0, \varepsilon)$. To see this, note that

$$\sin \tfrac{1}{x} = \begin{cases} 0 & \text{if } x = \frac{2}{2n\pi}, \quad n \in \mathbb{N} \\ 1 & \text{if } x = \frac{2}{(4n+1)\pi}, \quad n \in \mathbb{N} \\ -1 & \text{if } x = \frac{2}{(4n+3)\pi}, \quad n \in \mathbb{N} \end{cases}$$

Thus if $x = \frac{2}{2\pi}, \frac{2}{3\pi}, \frac{2}{4\pi}, \frac{2}{5\pi}, \frac{2}{6\pi}, \ldots$, then $\sin \frac{1}{x} = 0, -1, 0, 1, 0, \ldots$. Now draw the graph of $\sin \frac{1}{x}$.

Note that $f(x)$ is not defined at $x = 0$. There is also no way that we can define $f(0)$ in such a way as to make f continuous at zero. For if we could define $f(0)$ to make f continuous, then we would have

$$f(0) = \lim_{x \to 0} \sin \frac{1}{x}$$

But we can find two sequences $x_n \to 0$ and $y_n \to 0$ such that $f(x_n) \to 1$ and $f(y_n) \to -1$. It follows that $\lim_{x \to 0} \sin \frac{1}{x}$ does not exist.

Next, consider the function

$$g(x) = x \sin \frac{1}{x}$$

which is also defined for $x \neq 0$. In this case, however, we *can* define $g(0)$ in such a way as to make g continuous: Simply put $g(0) = 0$. It is clear that whereas $f(x)$ bounces between -1 and 1, $g(x)$ bounces between $-x$ and x. A simple sandwich argument proves that g is continuous at 0:

$$-|x| \leq |x \sin \frac{1}{x}| \leq |x|$$

and thus $\lim_{x \to 0} x \sin \frac{1}{x} = 0$.

However, though g is continuous at $x = 0$, g is not differentiable there: For

$$g'(0) = \lim_{x \to 0} \frac{x \sin \frac{1}{x} - 0}{x - 0} = \lim_{x \to 0} \sin \frac{1}{x}$$

and $\lim_{x \to 0} \sin \frac{1}{x}$ does not exist. Thus $g'(0)$ does not exist either.

Next, consider the function

$$h(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

A simple sandwich argument shows that h is continuous at $x = 0$. h is also differentiable there:

$$h'(0) = \lim_{x \to 0} \frac{x^2 \sin \frac{1}{x} - 0}{x - 0} = \lim_{x \to 0} x \sin \frac{1}{x} = 0$$

However, the derivative h' is not continuous at $x = 0$:

$$\lim_{x \to 0} h'(x) = \lim_{x \to 0} \left[2x \sin \frac{1}{x} - \frac{x^2}{x^2} \cos \frac{1}{x} \right] = 0 - \lim_{x \to 0} \cos \frac{1}{x}$$

It is easy to show that $\lim_{x\to 0} \cos\frac{1}{x}$ does not exist, for the same reason that $\lim_{x\to 0} \sin\frac{1}{x}$ does not exist. Hence $\lim_{x\to 0} h'(x)$ does not exist, and thus it can certainly not be equal to $h'(0)$. This shows that h' is not continuous at $x = 0$.

Since a differentiable function is necessarily continuous, $h''(0)$ does not exist.

□

5.2 Mean Value Theorems

In this section, we continue to repay our debt to first year calculus, finally providing rigorous proofs for theorems that were just on loan then.[1] Thus we state and prove several well–known results, without probing for applications.

Definition 5.2.1 Let $f : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}$. We say that f has a *local maximum* at $x_0 \in X$ if and only if there is a neighbourhood U of x_0 such that $f(x) \leq f(x_0)$ for all $x \in U$.
f has a local minimum at x_0 if $-f$ has a local maximum at x_0.

Remarks 5.2.2 It is easy to see that f has a local maximum at x_0 if and only if there exists a $\delta > 0$ such that

$$f(x) \leq f(x_0) \quad \text{whenever} \quad |x - x_0| < \delta \qquad (x \in X)$$

□

Perhaps the single most useful fact in differential calculus is the following:

Theorem 5.2.3 *Suppose X is an interval, and that x_0 is an interior point of X. If a function $f : X \to \mathbb{R}$ has a local maximum at x_0, and if $f'(x_0)$ exists, then $f'(x_0) = 0$.*

Proof: Choose $\delta > 0$ such that

$$f(x) \leq f(x_0) \quad \text{whenever} \quad |x - x_0| < \delta \qquad (x \in X)$$

and such that also $(x_0 - \delta, x_0 + \delta) \subseteq X$ (which can be done because x_0 is assumed to be an interior point of X). Now if $x_0 - \delta < x < x_0$, we have

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0$$

because $f(x) - f(x_0)$ and $x - x_0$ are both negative. On the other hand, if $x_0 < x < x_0 + \delta$, then

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0$$

because $f(x) - f(x_0)$ is negative, but $x - x_0$ is positive. Hence if $\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x-0}$ exists, then it is both ≥ 0 and ≤ 0. The result follows.

⊣

[1] I'll bet that the rate of interest is very low...

Theorem 5.2.4 (Rolle's Theorem)
Suppose that $f : [a, b] \to \mathbb{R}$ is continuous, with $f(a) = f(b) = 0$. Suppose further that that f is differentiable on the open interval (a, b). Then there exists a $c \in (a, b)$ such that $f'(c) = 0$.

Proof: This is obvious if $f = 0$ on $[a, b]$. Suppose therefore, that f is not identically zero. Thus there exists an $x \in (a, b)$ such that $f(x) \neq 0$. Replacing f by $-f$, if necessary, we may assume that f takes on some strictly positive value.
Now $[a, b]$ is a compact interval, and f is continuous. Hence f attains its supremum, i.e. there is $c \in [a, b]$ such that $f(c) = \sup_{a \leq x \leq b} f(x)$. Since this supremum is > 0, and since $f(a) = 0 = f(b)$, we see that in fact $c \in (a, b)$. Thus $f'(c)$ exists, and, by Theorem 5.2.3, we have $f'(c) = 0$, as required.

$\dashv$

The next theorem states that if $f : [a, b] \to \mathbb{R}$ is differentiable, there is some point c between a and b such that the slope of the tangent at point c is equal to the slope of the chord joining the points $(a, f(a))$ and $(b, f(b))$ on the graph of f. This is a central result in differential calculus:

Theorem 5.2.5 (Mean Value Theorem)
Suppose that $f : [a, b] \to \mathbb{R}$ is continuous, and that f is differentiable on the open interval (a, b). Then there exists a point $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)(b - a)$$

In fact, we can do better:

Theorem 5.2.6 (Generalized Mean Value Theorem)
Suppose that f, g are continuous on the closed interval $[a, b]$, and differentiable on the open interval (a, b). Then there exists a point $c \in (a, b)$ such that

$$[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c)$$

This result is also called the *Cauchy Mean Value Theorem.* The proof is an exercise:

Exercise 5.2.7 (a) Prove the Generalized Mean Value Theorem for the case $g(b) \neq g(a)$ by applying Rolle's Theorem to the function

$$h(x) = f(x) - f(a) - \frac{f(b) - f(a)}{g(b) - g(a)}(g(x) - g(a))$$

Also prove the result in case $g(b) = g(a)$.

(b) Hence prove the (ordinary) Mean Value Theorem.

$\square$

As an (important) application, we prove L'Hôpital's Rule. Here's a warm–up exercise:

Exercise 5.2.8 (1) Suppose that f, g are on $[a, b]$ and differentiable on (a, b). Suppose further that $f(a) = 0 = g(a)$, and that g, g' do not vanish on (a, b). We prove that

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}$$

(a) Choose x such that $a < x < b$. Use the Generalized Mean Value Theorem to prove that there is a $c \in (a, x)$ such that

$$\frac{f(x)}{g(x)} = \frac{f'(c)}{g'c)}$$

(b) Now let $x \to a$, and deduce the required result.

(2) Define

$$f(x) = \begin{cases} \dfrac{x-1}{\ln x} & \text{if } x \neq 1 \\ 1 & \text{if } x = 1 \end{cases}$$

Show that f is continuous at $x = 1$.

□

Theorem 5.2.9 (L'Hôpital's Rule)
Suppose that f, g are differentiable on (a, b), where $-\infty \leq a < b \leq \infty$, and that $g'(x) \neq 0$ for $x \in (a, b)$. Suppose further that

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$$

Then if either

(a) $\lim_{x\to a} f(x) = 0 = \lim_{x\to a} g(x)$, *or*

(b) $\lim_{x\to a} g(x) = \infty$,

then also

$$\lim_{x\to a} \frac{f(x)}{g(x)} = L$$

A similar result holds if $x \to b$ or $g(x) \to -\infty$.

Proof: Step I: First suppose that $-\infty < L < +\infty$.
(a) Now assume also that $\lim_{x\to a} = 0 = \lim_{x\to a} g(x)$. Let $\varepsilon > 0$. Since $\frac{f'(x)}{g'(x)} \to L$ as $x \to a$, there exists a number $c \in (a, b)$ such that

$$\frac{f'(x)}{g'(x)} \leq L + \varepsilon \qquad \text{whenever} \qquad a < x < c$$

Then if $a < x < y < c$, we see that there is a $t \in (x, y)$ such that

$$g'(t)[f(x) - f(y)] = f'(t)[g(x) - g(y)]$$

by the Generalized Mean Value Theorem. It follows that

$$\frac{f(x)-f(y)}{g(x)-g(y)} = \frac{f'(t)}{g'(t)} < L+\varepsilon$$

(because $a < t < c$) and thus

$$\frac{f(x)-f(y)}{g(x)-g(y)} < L+\varepsilon$$

Now let $x \to a$ to obtain $\frac{f(y)}{g(y)} \leq L+\varepsilon$. Thus we have shown that there is a number $c > a$ such that

$$\frac{f(y)}{g(y)} \leq L+\varepsilon \qquad \text{whenever} \qquad a < y < c \tag{$*$}$$

(b) Now assume that $\lim_{x\to a} g(x) = +\infty$. As above, we can find a $c' > a$ such that

$$\frac{f'(x)}{g'(x)} < L+\varepsilon \qquad \text{whenever } a < x < c'$$

and thus, by a similar application of the Mean Value Theorem, that

$$\frac{f(x)-f(y)}{g(x)-g(y)} < L+\varepsilon \qquad \text{whenever } a < x < y < c'$$

Now fix $y \in (a, c')$. Because $g(x) \to \infty$ as $x \to a$, we can find a c'' such that $a < c'' < y$ such that

$$g(x) > g(y) \qquad \text{and} \qquad g(x) > 0 \quad \text{for all } x \in (a, c)$$

Then

$$\frac{f(x)-f(y)}{g(x)-g(y)} = \frac{f'(t)}{g'(t)} < L+\varepsilon$$

Multiply both sides of this equation by $\frac{g(x)-g(y)}{g(x)}$ to obtain, after some rearranging,

$$\frac{f(x)}{g(x)} < (L+\varepsilon)\left(1-\frac{g(y)}{g(x)}\right) + \frac{f(y)}{g(x)} \qquad \text{for all } x \in (a, c'')$$

Now as $x \to a$, $g(x) \to \infty$, and thus $\frac{g(y)}{g(x)} \to 0$, $\frac{f(y)}{g(x)} \to 0$. It follows that $\frac{f(x)}{g(x)} \leq L+\varepsilon$ for x sufficiently close to a, i.e. there is $c > a$ such that

$$\frac{f(x)}{g(x)} \leq L+\varepsilon \qquad \text{whenever} \qquad a < x < c \tag{$**$}$$

Comparing $(*)$ and $(**)$, we see that in both cases (a) and (b) we have shown that there exists a $c > a$ such hat $\frac{f(x)}{g(x)} \leq L+\varepsilon$ whenever $a < x < c$.

A similar argument shows that, for both cases (a) and (b), we can find c' such that

$$\frac{f(x)}{g(x)} \geq L-\varepsilon \qquad \text{whenever} \qquad a < x < c'$$

Now combine these results: Given any $\varepsilon > 0$, we can find a c such that $L-\varepsilon \leq \frac{f(x)}{g(x)} \leq L+\varepsilon$ whenever $x \in (a, c)$. This is easily seen to imply that $\frac{f(x)}{g(x)} \to L$.

Step II: Now assume that $L = -\infty$.
Further assume that $\lim_{x\to a} f(x) = 0 = \lim_{x\to a} g(x)$. Let K be any real number. Arguing as in Step I(a), we can find $c > a$ such that $\frac{f(x)}{g(x)} \leq K$ whenever $a < x < c$. (Just replace $L + \varepsilon$ by K.) Arguing as in Step I(b), we can conclude the same thing if $\lim_{x\to a} g(x) = +\infty$. Since K is arbitrary, it follows that $\frac{f(x)}{g(x)} \to -\infty = L$ as $x \to a$.

A similar argument works if $L = +\infty$.

⊣

Exercise 5.2.10 Fill in the missing arguments in Step II of the proof of L'Hôpital's Rule.

□

Exercise 5.2.11 Discuss the following argument:

Statement: *If $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then the derivative function f' is continuous.*
Proof:

(i) To prove that a function $u : \mathbb{R} \to \mathbb{R}$ is continuous at a point x, it suffices to show that $\lim_{h\to 0} u(x + h) = u(x)$

(ii) Suppose that f is differentiable at x. Define $g(h) = \frac{f(x+h)-f(x)}{h}$. Then $f'(x) = \lim_{h\to 0} g(h)$.

(iii) By L'Hôpital's rule and the chain rule

$$f'(x) = \lim_{h\to 0} g(h) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{f'(x+h)\cdot 1 - 0}{1} = \lim_{h\to 0} f'(x+h)$$

(iv) Since $\lim_{h\to 0} f'(x + h) = f'(x)$, it follows that f' is continuous at x.

(v) Since x was an arbitrary point, f' is a continuous function.

□

Exercise 5.2.12 Define

$$f(x) = x + \cos x \sin x \qquad g(x) = e^{\sin x} f(x)$$

(a) Show that $\lim_{x\to\infty} f(x) = +\infty = \lim_{x\to\infty} g(x)$

(b) Show that $\lim_{x\to\infty} \frac{f(x)}{g(x)}$ does not exist.

(c) Show that $f'(x) = 2(\cos x)^2$ and $g'(x) = e^{\sin x} \cos x[2\cos x + f(x)]$

(d) Show that $\frac{f'(x)}{g'(x)} = \frac{2e^{-\sin x}\cos x}{2\cos x + f(x)}$ if $\cos x \neq 0$

(e) Conclude that $\lim_{x\to\infty} \frac{f'(x)}{g'(x)} = 0$.

Here we have a situation where f, g are differentiable functions with $\lim_{x\to\infty} f(x) = \infty = \lim_{x\to\infty} f(x)$, yet l'Hôpital's rule doesn't seem to work! The reason is as follows: We saw that $\dfrac{f'(x)}{g'(x)} = \dfrac{2e^{-\sin x}\cos x}{2\cos x + f(x)}$ if $\cos x \neq 0$. However, as $x \to \infty$, $\cos x = 0$ for infinitely many x. It is therefore not the case that $\dfrac{f'(x)}{g'(x)} = \dfrac{2e^{-\sin x}\cos x}{2\cos x + f(x)}$ for all x.

For $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$ to exist, it is necessary that $g'(x)$ is non–zero for all sufficiently large x. Similarly, for $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ to exist, it is necessary that $g'(x)$ is non–zero in some neighbourhood of a.

□

The final result in this section is Taylor's Theorem, which plays an important role in numerical analysis. Recall that $f^{(k)}$ denotes the k^{th} derivative of f. Also, let $f^{(0)} = f$. Finally, recall that $0! = 1$.

Theorem 5.2.13 (Taylor's Theorem)
Let $n \in \mathbb{N}$. Suppose that $f : [a, b] \to \mathbb{R}$ is a function with the property that

(i) $f, f', f'', \ldots, f^{(n-1)}$ are defined and continuous on $[a, b]$;

(ii) $f^{(n)}$ exists on (a, b).

Suppose further that α, β are real numbers such that $a \leq \alpha < \beta \leq b$. Then there exists a $\gamma \in (\alpha, \beta)$ such that

$$f(\beta) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(\beta - \alpha)^k + \frac{f^{(n)}(\gamma)}{n!}(\beta - \alpha)^n$$

Proof: For $x \in [a, b]$, define

$$P(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(x - \alpha)^k$$

We must show that there is a $\gamma \in (\alpha, \beta)$ such that $f(\beta) = P(\beta) + \frac{f^{(n)}(\gamma)}{n!}(\beta - \alpha)^n$. Now let $M = \frac{f(\beta) - P(\beta)}{(\beta - \alpha)^n}$, so that

$$f(\beta) = P(\beta) + M(\beta - \alpha)^n$$

and define

$$g(x) = f(x) - P(x) - M(x - \alpha)^n \qquad (x \in [a, b])$$

Note that $g(\beta) = 0$, by definition of M. Note also that

$$g^{(k)}(\alpha) = f^{(k)}(\alpha) - P^{(k)}(\alpha) = 0 \qquad k = 0, 1, \ldots, n-1$$

because $P^{(k)}(\alpha) = f^{(k)}(\alpha)$ for such k.

We must show that $M = \frac{f^{(n)}(\gamma)}{n!}$ for some γ between α and β. Note that $P(x)$ is an $(n-1)^{\text{th}}$ degree polynomial in x, so that $P^{(n)}(x) = 0$. It follows that

$$g^{(n)}(x) = f^{(n)}(x) - n!M$$

Thus if we can find a γ such that $g^{(n)}(\gamma) = 0$, we will have $M = \frac{f^{(n)}(\gamma)}{n!}$, as required.

Now both $g(\alpha) = 0$ and $g(\beta) = 0$. By the Mean Value Theorem, there is $\gamma_1 \in (\alpha, \beta)$ such that $g'(\gamma_1) = 0$. Thus both $g'(\alpha) = 0$ and $g'(\gamma_1) = 0$. By the Mean Value Theorem, there is $\gamma_2 \in (\alpha, \gamma_1)$ such that $g''(\gamma_2) = 0$. Thus both $g''(\alpha) = 0$ and $g''(\gamma_2) = 0$. By the Mean Value Theorem, there is $\gamma_3 \ldots$

After $n-1$ steps, we obtain, from the fact that both $g^{(n-1)}(\alpha) = 0$ and $g^{(n-1)}(\gamma_{n-1}) = 0$, a $\gamma_n \in (\alpha, \gamma_{n-1})$ such that $g^{(n)}(\gamma_n) = 0$.

γ_n is therefore the γ that we seek.

⊣

Remarks 5.2.14 Note that if $n = 1$, Taylor's Theorem is just the ordinary Mean Value Theorem.

□

If f is a function with derivatives of all orders, then the *Taylor series* of f about a point $x = a$ is defined by

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

We will have more to say about Taylor *series* in a future chapter. For the moment, however, let's just show that the well–known Taylor series for e^x converges, and to the right value, for each $x \in \mathbb{R}$:

Exercise 5.2.15

Let f be the function $f(x) := e^x$. Fix $x \neq 0$, and let $[a, b]$ be an interval containing both 0 and x. The aim of this problem is to show that the power series

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \qquad (\dagger)$$

converges to e^x.

5.1 Show that the power series (†) converges, irrespective of the value of x.

5.2 Use Taylor's Theorem, as stated above, to show that

$$e^x = \sum_{k=0}^{n} \frac{x^k}{k!} + R_{n+1}(x)$$

where

$$|R_{n+1}(x)| \leq C\frac{x^{n+1}}{(n+1)!} \qquad C \text{ is a constant}$$

5.3 Show that $\lim_{n\to\infty} R_{n+1}(x) = 0$, irrespective of the value of x.

5.4 Now explain why $\sum_{k=0}^{\infty} \frac{x^k}{k!}$ converges to e^x.

□

It is unfortunately not the case that the Taylor series of a function always converges to the correct value, however. Here is an example:

Example 5.2.16 Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) := \begin{cases} e^{-\frac{1}{x^2}} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

We claim that the Taylor series for $f(x)$ is $\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = 0$, because $f^{(n)}(0) = 0$ for all n. Thus the Taylor series does converge for every x, but it never converges to the correct value e^{-1/x^2}, except at $x = 0$.

It is not hard to see that f is continuous, i.e. that $\lim_{x\to 0} f(x) = 0$. Note that if $x \neq 0$, then

$$f'(x) = \tfrac{2}{x^3} f(x) \qquad f''(x) = (\tfrac{4}{x^6} - \tfrac{6}{x^4}) f(x)$$

etc., and that, generally

$$f^{(n)}(x) = P_n(\tfrac{1}{x}) f(x) \qquad x \neq 0$$

where $P_n(x)$ is a polynomial. Now by L'Hôpital's Rule,

$$\lim_{x\to 0} \frac{f(x)}{x} = \lim_{x\to 0} \frac{e^{-1/x^2}}{x} = \lim_{x\to 0} \frac{1/x}{e^{1/x^2}} = 0$$

Assume now that $\lim_{x\to 0} \frac{f(x)}{x^k} = 0$ holds for $k = 0, 1, \ldots, n$. Then, again by L'Hôpital's Rule,

$$\lim_{x\to 0} \frac{f(x)}{x^{n+1}} = \lim_{x\to 0} \frac{x^{-(n+1)}}{e^{1/x^2}} = \frac{n+1}{2} \lim_{x\to 0} \frac{f(x)}{x^{n-1}} = 0$$

By induction, and taking linear combinations, it follows that

$$\lim_{x\to 0} P(\tfrac{1}{x}) f(x) = 0 \qquad \text{for any polynomial } P$$

Now

$$f'(0) = \lim_{h\to 0} \frac{f(h)}{h} = 0$$

Assume now that $f^{(k)}(0) = 0$ for $k = 1, \ldots, n$. Then

$$\begin{aligned} f^{(n+1)}(0) &= \lim_{h\to 0} \frac{f^{(n)}(h) - f^{(n)}(0)}{h} \\ &= \lim_{h\to 0} \frac{P_n(1/h) f(h)}{h} \\ &= \lim_{h\to 0} Q_n(1/h) f(h) \\ &= 0 \end{aligned}$$

because $Q_n(x) := x P_n(x)$ is a polynomial. By induction, $f^{(n)}(0) = 0$ for all n.

□

Appendix A

Logic, Sets and Functions

A.1 Logic and Formal Language

We introduce here a *formal language* for talking about mathematical objects. This language is very precise, and unambiguous — properties which are largely absent from spoken languages such as English, but obviously essential for mathematics. But, as a result, this language is rather *restricted* in scope. The reason we use it is to make certain statements amenable to *logical analysis*. The purpose of logical analysis is to decide whether a particular sentence/expression (e.g. about mathematical objects) is *true* (T) or *false* (F). A sentence/expression that is either true or false (but not both!) is called a **statement**.

Example A.1.1 Here are some typical examples of statements:

- $1 + 1 = 3$.
- All apples are red.
- The equation $x^2 + 2x + 1 = 0$ has a real root.
- Either $x^2 + a = 0$ has a real root, or $a > 0$.
- There exist are infinitely many prime numbers.
- Every continuous function is differentiable.

Note that a mathematical statement need not be true.

□

Exercise A.1.2 Which of the following are statements? For each statement, try to decide whether it is true or false.

1. $1 + 1$
2. 3 is greater than 0.
3. $\sqrt{2}$ is an irrational number.
4. $x^2 - 1 = 0$.

5. If $x = 1$, then $x^2 - 1 = 0$.
6. If $x^2 - 1 = 0$, then $x = 1$.
7. The moon is made of cheese.
8. The moon is a tasty snack.
9. If the moon is made of cheese, then the moon is a tasty snack.
10. The sentence ϕ defined by

$$\phi \equiv \text{“The sentence } \phi \text{ is false”}$$

11. All unicorns are white.
12. All unicorns are pink.

□

More complicated statements in our *formal language* are built up from a collection of symbols, including amongst others

- Symbols for objects and relations;
- Logical Connectives;
- Quantifiers;

We will briefly discuss each of these in turn. None of this material is difficult, though it may take a little while to get used to.

A.1.1 Symbols denoting Objects, Operations and Relations

When doing mathematics, we use symbols to denote certain mathematical objects, operations and relations. For example, the expression

$$x + 3 \leq \sqrt{\pi}$$

contains the following symbols:

(i) Symbols denoting fixed objects, namely the constants 3 and π;

(ii) A symbol denoting a variable object, namely x;

(iii) Symbols denoting operations, namely $+, \sqrt{}$;

(iv) A symbol denoting a relationship, namely $\leq$;

So our language will contain symbols for

- **Variables:** Typically we use the symbols $x, y, z, x_1, x_2, x_3 \ldots$
- **Constants:** e.g. $0, 1, 2, \ldots$, or π, etc.
- **Functions/Operations:** $+, \cdot, \sqrt{}, \cup, \cap$
- **Properties and Relations:** e.g. $=, \leq, >, \in, \subseteq$, etc.

A.1.2 Logical Connectives

Once we are able to make basic statements such as $1 > 0$ and $x = 3$, we are able to combine them using the logical connectives *and, or, implies (then), not* to make new statements such as

$$(1 > 0) \text{ and } (x = 3); \qquad \text{If } x > 0 \text{ then } y = 1; \qquad x \not\geq 0$$

$\wedge$	and
$\neg$	not
$\rightarrow$	implies, then
$\vee$	or
$\leftrightarrow$	if and only if

In our formal language, these connectives have precise meanings: If ϕ, ψ denote statements, then

$\phi \wedge \psi$ is true	$\Longleftrightarrow$	*both* ϕ, ψ are true.
$\phi \vee \psi$ is true	$\Longleftrightarrow$	at least one of ϕ, ψ is true, perhaps both.
$\phi \rightarrow \psi$ is true	$\Longleftrightarrow$	whenever ϕ is true, so is ψ
		i.e. it is not the case that ϕ is true but ψ is false.
$\phi \leftrightarrow \psi$ is true	$\Longleftrightarrow$	if *both* $\phi \rightarrow \psi$, $\psi \rightarrow \phi$ are true
		i.e. if ϕ, ψ are simultaneously true, or when they are simultaneously false.
$\neg\phi$ is true	$\Longleftrightarrow$	ϕ is false.

Here is a *truth table* for the logical connectives:

ϕ	ψ	$\phi \wedge \psi$	$\phi \vee \psi$	$\phi \rightarrow \psi$	$\phi \leftrightarrow \psi$	$\neg\phi$
T	T	T	T	T	T	F
T	F	F	T	F	F	F
F	T	F	T	T	F	T
F	F	F	F	T	T	T

This means, for example, that if ϕ is true and ψ is false — in the second row of the table — then $\phi \wedge \psi$ is false, $\phi \vee \psi$ is true, $\phi \rightarrow \psi$ is true, etc.

Now it is **extremely important** to note that the logical use of *and* $\wedge$, *or* $\vee$, and *implies* $\rightarrow$, though related to their common usage in English, is certainly not identical to it. In particular the *truth value* T or F of an expression such as $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$ etc. depends only on the truth values of ϕ and ψ, and *not* on any meaning that the statements ϕ, ψ might possess! Let us discuss some of the pitfalls:

- **And, $\wedge$:**

ϕ	ψ	$\phi \wedge \psi$
T	T	T
T	F	F
F	T	F
F	F	F

  To say that $\psi \wedge \psi$ simply means that both ϕ *and* ψ are true. It does not assert any connection (causal or otherwise) between ϕ and ψ. This is not typically true in English. With the English *and*, the following sentences have rather different meanings, but with the logical *and* they mean the same thing:

1. Alice got drunk and failed her test.
2. Alice failed her test and got drunk.

- **Or, $\vee$:**

ϕ	ψ	$\phi \vee \psi$
T	T	T
T	F	T
F	T	T
F	F	F

$\phi \vee \psi$ is true precisely when at least one of ϕ, ψ is true, possibly both. In particular, it is not *exclusive–or* ("either..., or..."). Thus the statement

$$(1 > 0) \vee (5 \text{ is a prime number})$$

is true.

- **Implies, Then, $\rightarrow$:**

ϕ	ψ	$\phi \vee \psi$
T	T	T
T	F	F
F	T	T
F	F	T

The statement $\phi \rightarrow \psi$ is true if whenever ϕ is true, then so is ψ. In particular,

$$\phi \rightarrow \psi \text{ is false if and only if } \phi \text{ is true but } \psi \text{ is false.}$$

There are *severe* differences between the English usage and the mathematical usage of *implies*. In English usage, *implies* (or *then*) usually involves a causal connection, as in "If it is raining, then it is wet outside." It is wet *because* of the rain. But such a connection is irrelevant for the logical *then*. For example, the statement

$$(1 > 0) \rightarrow (5 \text{ is a prime number})$$

is true. Of course, the reason that 5 is prime is not because of the fact that $1 > 0$!! There is no causal connection.

We repeat: A logical $\phi \rightarrow \psi$ statement is false *only* when ϕ is true and ψ is false — just look at the truth table.

- In particular, if ψ is true, then $\phi \rightarrow \psi$ is also true, no matter what ϕ might be.
- Even more surprisingly, if ϕ is false, then $\phi \rightarrow \psi$ is true, i.e. *a false statement implies any other statement*! In particular

$$(0 = 1) \rightarrow (\text{The Moon is made of cheese})$$

is true.

Exercise A.1.3 (a) Two statements P, Q are said to be *logically equivalent* — and we write this as $P \Longleftrightarrow Q$ — if and only if P, Q have the same truth value. There is an algorithm

to check if two statements are logically equivalent: Simply construct a truth table for P, Q and show that the truth values for P, Q are always the same.[1] Show that

$$\neg\neg\phi \iff \phi$$

that

$$\neg(\phi \vee \psi) \iff (\neg\phi) \wedge (\neg\psi)$$

$$\neg\neg\phi \iff \phi$$

that

$$\neg(\phi \vee \psi) \iff (\neg\phi) \wedge (\neg\psi)$$

and that

$$\phi \to \psi \iff (\neg\phi) \vee \psi$$

[Hint: For the first equivalence, construct the truth table

ϕ	ψ	$\phi \vee \psi$	$\neg(\phi \vee \psi)$	$\neg\phi$	$\neg\psi$	$(\neg\phi) \wedge (\neg\psi)$
T	T					
T	F					
F	T					
F	F					

The truth value entries in the $\neg(\phi \vee \psi)$–column and the $(\neg\phi) \wedge (\neg\psi)$–column should be identical (or else you've made a mistake). This means that $\neg(\phi \vee \psi)$ is true precisely when $(\neg\phi) \wedge (\neg\psi)$ is true, and hence that the statements are equivalent. Repeat for the other equivalences that must be shown.]

(b) Show that

$$(\phi \to \psi) \iff (\neg\psi) \to (\neg\phi)$$

This is important in proofs: To show that ψ follows from ϕ it is enough to show that if ψ fails to be true, then ϕ also fails to be true.

□

Exercise A.1.4 A **proof by contradiction** works as follows: To prove that a statement P is true, it is enough to show that there is a *known false statement* Q so that $\neg P \to Q$ is true, i.e. so that assuming that P is not true leads to a false statement. We may then conclude that P is true (for if P were false, then $\neg P$ would be true, and since $\neg P \to Q$ is true, we may conclude that Q is true — contradicting the fact that Q is known to be false.)

Can you demonstrate the above reasoning using a truth table?
[Hint: Construct a truth table as above, and then remove rows that contradict what you know. You know Q is false, so you can remove rows in which Q is true. You know $\neg P \to Q$ is true, so . . .]

□

[1] This method probably appears first in Ludwig Wittgenstein's *Tractatus Logico–Philosophicus*, but he undoubtedly cribbed the idea from Gottlob Frege's *Begriffschrift*.

A.1.3 Quantifiers

Many mathematical statements assert the *existence* of a mathematical object with certain properties. For example to say that

$$x^2 - 1 = 0 \quad \text{has a real root}$$

is to say that there exists a real number c such that $c^2 - 1 = 0$.
Other mathematical statements assert that something is true for all objects (of a prespecified type), for example

For every real number x, $x^2 \geq 0$.

We therefore introduce the following symbols for *quantifiers*:

$\forall$	For all
$\exists$	There exists

A quantifier always occurs in conjunction with a variable, i.e. as $\forall x$ or as $\exists x$. Thus if $\phi(x)$ is a statement about x, then

$$\forall x \phi(x) \quad \text{is true iff the statement } \phi(x) \text{ is true for every } x$$

Frequently, if we want to restrict the domain to a particular set X, we may also write $\forall x \in X\ \phi(x)$ or $\exists x \in X\ \phi(x)$. Thus

$$(\exists x \in X)\phi(x) \quad \text{is true iff there is at least one } x \in X \text{ for which the statement } \phi(x) \text{ is true}$$

Thus the statement $\exists x \in \mathbb{R}(x^2 - 1 = 0)$ asserts that the equation $x^2 - 1 = 0$ has a real root.
The statement $\forall x \in \mathbb{R}(x^2 \geq 0)$ asserts that the square of any real number is non–negative.

Exercise A.1.5 Decide if the following sentences about real numbers are true or false:

(a) $\exists x \in \mathbb{R}(x^2 = -1)$

(b) $\exists x \in \mathbb{N}(4x = 1)$

(c) $\exists x \in \mathbb{R}(4x = 1)$

(d) $\forall x \in \mathbb{R}\ \exists y \in \mathbb{R}(x \leq y)$

(e) $\exists y \in \mathbb{R}\ \forall x \in \mathbb{R}(x \leq y)$

(f) $\exists y \in [0,1]\ \forall x \in [0,1](x \leq y)$

(g) $\forall x \in \mathbb{R}\ \forall y \in \mathbb{R}[xy = 0 \rightarrow (x = 0 \vee y = 0)]$

(h) $\forall x \in \mathbb{R}\ \forall y \in \mathbb{R}\ \exists z \in \mathbb{R}[x + z = y]$

(i) $\exists z \in \mathbb{R}\ \forall x \in \mathbb{R}\ \forall y \in \mathbb{R}[x + z = y]$

□

Exercise A.1.6 Rewrite the following sentences about numbers using logical notation.

(a) The integer x is an even number. [Hint: An integer x is even if and only if there is an integer y such that $x = 2y$]

(b) x is an odd number.

(c) Any integer is either odd or even.

(d) For any positive integer, there is another integer so that their sum is negative.

(e) x is a rational number.

(f) $\sqrt{2}$ is an irrational number.

□

Note that we have the following equivalence of statements:

$$\neg\Big(\forall x\varphi(x)\Big) \quad \Longleftrightarrow \quad \exists x(\neg\varphi(x))$$

For if it isn't the case that the statement $\varphi(x)$ is true for every x, then there is at least one x for which the statement $\varphi(x)$ is false, and thus for which $\neg\varphi(x)$ is true.

Exercise A.1.7 Verduidelik waarom

$$\neg\Big(\exists x\varphi(x)\Big) \quad \Longleftrightarrow \quad \forall x(\neg\varphi)$$

□

Thus a negation sign can "creep" past a quantifier, but it *flips* the quantifier in the process. For example,

$$\begin{aligned}\neg[\forall x\exists y(y > x)] &\Longleftrightarrow \exists x\neg[\exists y(y > x)] \\ &\Longleftrightarrow \exists x\forall y(y \not> x)\end{aligned}$$

One more thing: The variable x in a statement of the form $\forall x\phi(x)$ or $\exists x\phi(x)$ is unimportant, i.e. the meaning of the statement remains the same if we change the variable (provided that the new variable does not already occur in the statement ϕ). This is just like what happened for definite integrals: For example, we have

$$\int_a^b f(x)\,dx = \int_a^b f(y)\,dy$$

Just so, we have

$$\forall x\phi(x) \quad \Longleftrightarrow \quad \forall y\phi(y) \qquad \text{and} \qquad \exists x\phi(x) \quad \Longleftrightarrow \quad \exists y\phi(y)$$

provided y does not already occur in ϕ.

A.2 Sets, Functions and Relations

The philosophical debate about the *nature* of mathematical objects was given a boost when it became generally accepted (in the early 20th century) that, in principle, *all* mathematical objects "should" be sets and mathematical notions "should" be expressible as relationships between sets. This means, for example, that $\sqrt{2}$ is a set!! Actually, you mustn't take this too literally — What is meant is that set theory is flexible enough to *interpret* all mathematical objects as sets. We have mentioned before that the first satisfactory answers to the question "What is a real number?" were given independently, but nearly simultaneously (1872) by Cantor and Dedekind:

- Cantor: A real number a is a certain set of sequences of rational numbers (namely the set of all such sequence that converge to a — but the definition can be phrased in such a way as to remove the circularity).
- Dedekind: A real number a is the a certain set of rational numbers (namely the set $\{x \in \mathbb{Q} : x < a\}$, but again the definition can be made non–circular).

The point is that both these approaches construct an object — a complete ordered field — that *behaves* just like the real numbers. The ingredients in the construction are simpler objects, namely rational numbers. Both approaches provide a *concrete construction* of an object that behaves just like the the set of reals. And we do not really care what real numbers *are*, but only how they *behave* and *interrelate*. [In the same way, a chess player doesn't care what a chess piece *is*. Whether a piece is made of wood, or plastic, or appears on a computer screen is completely irrelevant. What matters to the chess player is how the piece *behaves*, i.e. how the rules (axioms) allow it to interact with other pieces on the board.]

In the same way, almost any other mathematical object can be *interpreted* as a set, in some way or another. For this reason, every mathematician needs just a little set theory. The material in this section is not difficult, and no doubt you have seen it much of it before.

Intuitively, a *set* is just a collection of objects.

If A is a set and x is some mathematical object, we say that

$$x \in A \qquad (x \text{ is an } \textbf{element} \text{ of } A)$$

if x is amongst the objects collected in A, and we write

$$x \notin A$$

if it isn't.

The idea is that a set is *characterized entirely by its elements*. Thus if two sets A and B have exactly the same elements, then we must have $A = B$. For example, the sets $A = \{a\}$ and $B = \{a, a\}$ have the same elements, namely only a. Thus $A = B$. The fact that B seems to have two copies of a is immaterial.

Here are some remarks for the philosophically minded:

- Any definition involves some terms, and you can always ask for a definition of those terms. Those definitions will involve further terms, whose definition you can ask for... leading to either an infinite regress or circularity. We have to start *somewhere*, and we regard the notions of set an element as so basic that we need not define them: "We all know what is meant."

  And indeed, the idea of forming sets of objects *is* basic: If you want to count some objects, you first have to decide which objects you want to count, i.e. you first have to (mentally) put those objects in a set before you can count them. Forming sets is even more basic, therefore, than counting!!

It is absolutely remarkable that starting with just this undefined notion we can rigorously develop almost all mathematics, and certainly all applied mathematics.

- Saying that two sets are equal if and only if they have the same elements means, for example, that

$$\{\text{Evening Star}\} = \{\text{Morning Star}\}$$

as both sets are equal to the {planet *Venus*}. Yet the Evening Star is seen only in the evening, whereas the Morning Star is seen only in the morning...

Instead of *set*, we will also sometimes say *class*, *collection* or *family*; instead of saying x *is an element of* A we will sometimes say x *is a member of* A or x *belongs to* A.

There are two ways to represent sets:

(i) By *listing* its elements, and

(ii) By some defining *property*.

For example, if a set A has finitely many elements $a_1, \ldots, a_n$ then it can be represented by $A = \{a_1, a_2, \ldots, a_n\}$. On the other hand if A is the set of all x having a certain property $P(x)$, then A can be denoted by $A = \{x : P(x)\}$.

In analysis, the following sets are important:

- The set of natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$
- The set of integers or whole numbers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- The set of rational numbers $\mathbb{Q} = \{\frac{n}{m} : n, m \in \mathbb{Z}, m \neq 0\}$
- The set of real numbers $\mathbb{R}$, and the set of non–negative real numbers is denoted by $\mathbb{R}^+$.
- The set of complex numbers $\mathbb{C} = \{a + ib : a, b \in \mathbb{R}\}$

Example A.2.1

- The set A of all integers between -1 and 3 can be represented in two ways:

  (i) $A = \{-1, 0, 1, 2, 3\}$

  (ii) $A = \{n : n \text{ is an integer and } -1 \leq n \leq 3\}$

- $\mathbb{Q} = \{x \in \mathbb{R} : \exists n \in \mathbb{Z}\ (nx \in \mathbb{Z})\}$
- $\{\sqrt{2}\} = \{x \in \mathbb{R} : x > 0 \wedge x^2 = 2\}$.

□

A set need not have any elements:

Definition A.2.2 We define the *empty set* to be the set with no members, and denote it by the symbol $\emptyset$.

The empty set plays roughly the same role in set theory that the number zero plays in ordinary mathematics.

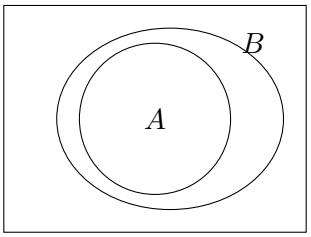


Figure A.1: Venn diagram illustrating *inclusion* $A \subseteq B$.

Exercise A.2.3 In the above definition, we speak of "the" emptyset. Explain why there is only one empty set, and not many. To be more concrete, note that both the sets $\{x : x \in \mathbb{R} \text{ and } x^2 < 0\}$ and $\{x : x \neq x\}$ have no elements. Explain why they are the same set.

□

Before we continue, please note the following common error:

$$A \neq \{A\}$$

e.g.

$$\emptyset \neq \{\emptyset\}$$

The set on the left has no elements, whereas the set on the right has one element, namely $\emptyset$.

Definition A.2.4 We say that a set A is a *subset* of another set B, and write

$$A \subseteq B$$

if and only if every element of A is also an element of B.
We say that A is a *proper subset* of B if A is subset of B, but $A \neq B$.

We may also write $B \supseteq A$ instead of $A \subseteq B$; they mean the same thing (just as $x \leq y$ and $y \geq x$ mean the same thing).

Remarks A.2.5 Note that $A = B$ if and only if $A \subseteq B$ *and* $B \subseteq A$.

□

Exercises A.2.6 (1) List all the subsets of the set $A := \{0, 1, 2, \{1, 2\}\}$,

(2) Prove that $\emptyset$ is a subset of every set.
[Hint: Give a proof by contradiction. Assume that there is a set A such that $\emptyset \not\subseteq A$.]

(3) Show formally that if $A \subseteq B$ and if $B \subseteq C$, then $A \subseteq C$.

(4) Prove (by induction or otherwise) that if a finite set A has n elements, then it has 2^n distinct subsets.

□

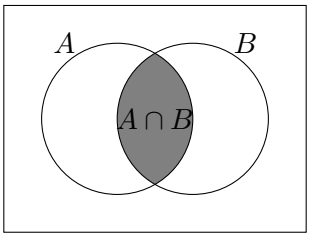


Figure A.2: Venn diagram illustrating *intersection* $A \cap B$.

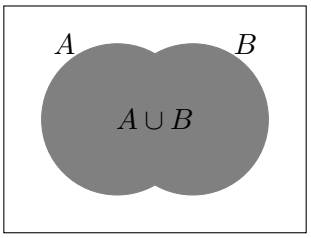


Figure A.3: Venn diagram illustrating *union* $A \cup B$.

A.2.1 Operations on sets

There are several ways of combining sets to form new sets. In this section we define and give some examples of the set–operations *union, intersection, difference, complementation, cartesian product* and *power set formation.*

Definition A.2.7 (Union, intersection and difference of two sets)
Suppose that A, B are sets.

(a) The *union* of A and B is the set of all elements which are either in A or in B (or both).
$$A \cup B = \{x : x \in A \vee x \in B\}$$

(b) The *intersection* of A and B is the set of all elements which belong to *both* A and B.
$$A \cap B = \{x : x \in A \wedge x \in B\}$$

(c) The *set difference* of A and B is the set of all elements which belong to A, but not to B.
$$A - B = \{x : x \in A \wedge x \notin B\}$$

Two sets A, B are said to be *disjoint* if they have no members in common, i.e. if $A \cap B = \emptyset$. In that case, $A - B = A, B - A = B$.

Often we work within some *universe*, which is just the set of all objects under consideration at that time. The sets that we deal with are then typically subsets of the universe.
Which set is the universe depends very much on context. If one is dealing with real numbers, the obvious choice of universe is $\mathbb{R}$, but if one is dealing with complex numbers as well, then it would be $\mathbb{C}$. If one is trying to find the solution of an n^{th} order differential equation, then the universe will generally be the set of all n–times differentiable functions. In probability theory, the sample space Ω acts as universe.

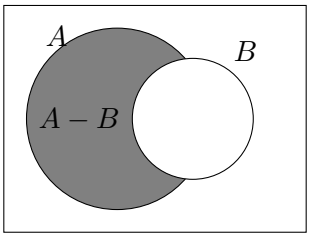


Figure A.4: Venn diagram illustrating *set difference* $A - B$.

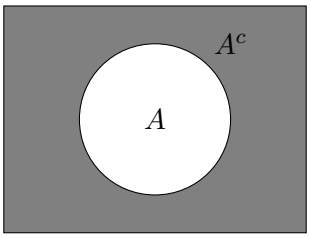


Figure A.5: Venn diagram illustrating *complementation* A^c.

Given a universe, we also have a unary operation on sets, called *complementation*.

Definition A.2.8 Let the universe be Ω, and let $A \subseteq \Omega$. The *complement* of A is the set of all elements in the universe which are not in A.

$$A^c = \{x \in \Omega : x \notin A\}$$

Note that $A^c = \Omega - A$. Also note that $A - B = A \cap B^c$.

Here are some standard identities involving the operations:

Proposition A.2.9 *Suppose that A, B, C are subsets of some universe Ω.*

(a) Idempotent laws:
$$A \cup A = A; \qquad A \cap A = A$$

(b) Commutative laws:
$$A \cup B = B \cup A; \qquad A \cap B = B \cap A$$

(c) Associative laws:
$$(A \cup B) \cup C = A \cup (B \cup C); \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

(d) Distributive laws:
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C); \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

(e) Absorption laws:
$$A \cup (A \cap B) = A; \qquad A \cap (A \cup B) = A$$

(f) Complementation laws:
$$A \cup A^c = \Omega; \qquad A \cap A^c = \emptyset$$
$$(A^c)^c = A$$

(g) De Morgan's laws:
$$(A \cap B)^c = A^c \cup B^c; \qquad (A \cup B)^c = A^c \cap B^c$$

Note that each of the identities remains true if

- $\cap$ and $\cup$ are interchanged, and
- $\emptyset$ and Ω are interchanged.

Proof: We show how to prove one of the above laws, and leave the remainder as exercises. Let us prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

First suppose that $x \in A \cap (B \cup C)$. Then $x \in A$ *and* $x \in B \cup C$, by definition of $\cap$. Thus $x \in A$ and *either* (1) $x \in B$, *or* (2) $x \in C$ (or both), by definition of $\cup$. Thus either (1) $x \in A$ and $x \in B$, *or* (2) $x \in A$ and $x \in C$. It follows that either (1) $x \in A \cap B$ *or* (2) $x \in A \cap C$, and thus that $x \in (A \cap B) \cup (A \cap C)$. We have now shown that if $x \in A \cap (B \cup C)$, then also $x \in (A \cap B) \cup (A \cap C)$, i.e. that

$$A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C) \qquad (*)$$

Next, assume that $x \in (A \cap B) \cup (A \cap C)$. Then either (1) $x \in A \cap B$, *or* (2) $x \in A \cap C$. In either case, it follows that $x \in A$. Also we must have either (1) $x \in B$, *or* (2) $x \in C$, and thus $x \in B \cup C$. We see, therefore, that we have *both* $x \in A$ *and* $x \in B \cup C$, so that $x \in A \cap (B \cup C)$. It follows that whenever $x \in (A \cap B) \cup (A \cap C)$, then also $x \in A \cap (B \cup C)$, i.e. that

$$(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C) \qquad (\dagger)$$

Putting $(*)$ and $(\dagger)$ together, we obtain

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

as required.

⊣

Exercise A.2.10 Prove the remaining identities in the proposition above.
(By the way, drawing a *Venn diagram* does **not** constitute a proof! Venn diagrams are drawings in the plane, and are reliable only when you are dealing with quite a small number of sets.)

□

A set is completely determined by its elements. The order in which those elements are arranged does not matter. For example, $\{a, b\} = \{b, a\}$. When we want the order to matter, we have to deal with ordered tuples. An *ordered pair* is denoted by (a, b), and should be thought of as a collection containing a and b, *in that order*. Thus $(a, b) \neq (b, a)$. Note that

$$(a, b) = (c, d) \iff a = c \text{ and } b = d$$

Generally, an *ordered n–tuple* is denoted by $(a_1, a_2, \ldots, a_n)$, and should be thought of as a collection containing $a_1, a_2, \ldots, a_n$, *in that order*.

The pair (a, b) is usually defined to be the *set* $\{\{a\}, \{a, b\}\}$. You can check that this definition yields the required property that $(a, b) = (c, d)$ iff $a = c$ and $b = d$.
(a, b, c) is then defined to be $(a, (b, c))$ (which is just the set $\{\{a\}, \{a, \{\{b\}, \{b, c\}\}\}\}$), etc. This is in keeping with the notion that all mathematical objects should be sets. On first encounter, however, you might find this arbitrary, clumsy, and unnecessary, and you wouldn't be far wrong: The *main* thing that you need to keep in mind is that *an ordered tuple is a collection in which the order matters.*

Using ordered tuples, we can define one more way of making new sets from old:

Definition A.2.11 (Cartesian product) Suppose that $A_1, A_2, \ldots, A_n$ are sets. The *cartesian product* of $A_1, \ldots, A_n$ is the set of *all* n–tuples $(a_1, \ldots, a_n)$, with each $a_k \in A_k$.

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_k \in A_k \text{ for } k = 1, 2, \ldots, n\}$$

We will identify the sets $(A \times B) \times C$ and $A \times (B \times C)$ with $A \times B \times C$, although, strictly speaking, they are not equal.

For example, $((a, b), c))$ is an element of the first set, but not of the second or third. $(a, (b, c))$ belongs to the second, but not to the first or third. (a, b, c) belongs to the third, but not to the first two. However, we shall simply *identify* $(a, (b, c)), ((a, b), c)$ and (a, b, c), i.e. we shall not distinguish between them. After all, all that matters is the order of a, b, c and that is the *same* in each of these tuples.

Thus far, we have considered union, intersection and cartesian product as *binary operations*, involving just two sets. Frequently, however, we may need to consider these as *infinitary* operations: We can, for example, take the union of infinitely many sets. We define the union, intersection and cartesian product of a family of sets as follows:

Definition A.2.12 (Union, intersection and product of a family of sets)
If $\mathcal{A} = \{A_i : i \in I\}$ is a family of sets, we may define

(a) the *union*

$$\bigcup \mathcal{A} = \bigcup_{i \in I} A_i = \{x : x \in A_i \text{ for some } i \in I\}$$

(b) the *intersection*

$$\bigcap \mathcal{A} = \bigcap_{i \in I} A_i = \{x : x \in A_i \text{ for all } i \in I\}$$

(c) the *cartesian product*

$$\prod \mathcal{A} = \prod_{i \in I} A_i = \{(a_i)_I : a_i \in A_i \text{ for all } i \in I\}$$

Here $(a_i)_I$ is a generalized tuple, indexed by I.
In essence, $(a_i)_I$ is a function with domain I and range $\bigcup_{i \in I} A_i$. We will return to this later.

We will frequently write $\bigcup_I A_i$ or $\bigcup_i A_i$ instead of $\bigcup_{i \in I} A_i$. We will also write $\bigcup_{n=1}^{\infty} A_n$ instead of $\bigcup_{n \in \mathbb{N}} A_n$. The same holds for $\bigcap$ and $\prod$.

Remarks A.2.13 Note that

(i) $\bigcup\{A, B\} = A \cup B$

(ii) $\prod\{A, B, C\} = A \times B \times C$

(iii) $\bigcap\{X_1, X_2, \ldots, X_n\} = X_1 \cap X_2 \cap \cdots \cap X_n$

etc.

□

Exercise A.2.14 Let $A_1, A_2, A_3, \ldots, A_n, \ldots$ be a sequence of subsets of a fixed set Ω. For $x \in \Omega$, we say that

$$x \in A_n \text{ eventually (ev.)}$$

if x belongs to all the A_n from some point onwards, i.e. if there exists an $N \in \mathbb{N}$ such that $x \in A_n$ for all $n \geq N$. (Then x belongs to all the A_n from N onwards.) Let (A_n, ev) denote the set of all x such that x belongs to A_n eventually, i.e.

$$(A_n, \text{ev.}) := \{x \in \Omega : x \in A_n, \text{ ev.}\}$$

Similarly, we say that

$$x \in A_n \text{ infinitely often (i.o.)}$$

if x belongs to infinitely many of the sets A_n, or, more accurately, if there are infinitely many $n \in \mathbb{N}$ such that $x \in A_n$. Let $(A_n, \text{i.o.})$ denote the set of all x such that x belongs to A_n infinitely often, i.e.

$$(A_n, \text{i.o.}) = \{x \in \Omega : x \in A_n, \text{i.o.}\}$$

(a) Explain why $(A_n, \text{ev.}) \subseteq (A_n, \text{i.o.})$.

(b) Explain why the following is true:

$$x \in A_n, \text{ev.} \quad \Longleftrightarrow \quad \exists N \in \mathbb{N} \forall n \geq N (x \in A_n)$$

(c) Explain why we may express $(A_n, \text{ev.})$ as follows:

$$(A_n, \text{ev.}) = \bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} A_n$$

[Hint: Try to understand the following reasoning: For $N \in \mathbb{N}$, define $B_N := \bigcap_{n \geq N} A_n = A_N \cap A_{N+1} \cap A_{N+2} \cap \ldots$. Then $x \in \bigcup_{N \in \mathbb{N}} \bigcap_{n \geq N} A_n$ iff $x \in \bigcup_{N \in \mathbb{N}} B_N$ iff there is some N such that $x \in B_N$.(Why?) Now $x \in B_N$ iff $x \in A_n$ for each $n \geq N$. (Why?)]

(d) Explain why the following is true:

$$x \in A_n, \text{i.o} \quad \Longleftrightarrow \quad \forall N \in \mathbb{N} \exists n \geq N (x \in A_n)$$

[Hint: If $x \in A_n$ i.o., then for each N, there must be $n \geq N$ such that $x \in A_N$. For if this were not so, then there would be some N such that $x \notin A_n$ for any $n \geq N$. But then x can belong only to those A_n for $n \in \{1, 2, \ldots, N-1\}$, i.e. to only finitely many of the A_n.]

(e) Explain why we may express $(A_n, \text{i.o.})$ as follows:

$$(A_n, \text{i.o.}) = \bigcap_{N \in \mathbb{N}} \bigcup_{n \geq N} A_n$$

(f) Explain why

$$(A_n, \text{ev.})^c = (A_n^c, \text{i.o.}) \qquad \text{and} \qquad (A_n, \text{i.o.})^c = (A_n^c, \text{ev.})^c$$

Do this in two ways: via logic, and via set theory.

□

Here is another way of making new sets from old: Given a particular set, one should be able to collect all of its subsets together into a new set, called the *power set*.

Definition A.2.15 (Power set)
If A is a set, then the *power set* of A is the set of all subsets of A.

$$\mathcal{P}(A) = \{B : B \subseteq A\}$$

Note that $\emptyset, A \in \mathcal{P}(A)$. They are, respectively, its smallest and biggest members.

A.2.2 Functions

Originally, a function was regarded as a *rule* (or a *formula*, or an *algorithm*) for associating one real number with another. For example,

$$f(x) = 2x^3$$

explicitly shows how to calculate a number $f(x)$ which is to be associated with x: First cube x, and then multiply the resultant by 2. However, this original formulation proved to be unduly restrictive. For one thing, Fourier showed that practically any continuous curve of finite length could be give a "formula" as an infinite trigonometric series. For another, we may want to associate numbers with other mathematical objects, or one kind of mathematical object with another — there is no reason to restrict ourselves solely to numbers.

For example, we may want to associate with each rectangle its area. Thus we have a function which assigns a number to each rectangle.
Or, we may want to assign to each subset of $\mathbb{R}$ its power set. This yields a function which assigns a set to each set.

Thus a general definition of function dispenses with the idea that it is a rule, but keeps the idea of associating one object with another:

Definition A.2.16 Let A, B be sets. A *function* (or *map*) f from A to B, written

$$f : A \to B \quad \text{or} \quad A \xrightarrow{f} B$$

is a subset of the cartesian product $A \times B$ with the following property:

$$\text{for each } a \in A \text{ there exists } \textit{exactly one } b \in B \text{ such that } (a, b) \in f$$

In that case write

$$f(a) = b \quad \text{instead of } (a, b) \in f$$

We call b the *image* (or *value*) of a under f, and call a a *preimage* of b. We also say that a *maps to* b under f.
The set A is called the *domain* of f, and the set B is called the *codomain* of f

$$A = \text{dom}(f) \qquad B = \text{codom}(f)$$

The *range* of f is the set of all possible values of f, and denoted $\text{ran}(f)$.

Essentially, this concept of function is arrived at by deliberately confusing a function with its graph. For example, the graph of the function $f : \mathbb{R} \to \mathbb{R} : x \mapsto 2x^3$ is a curve in the cartesian plane. This curve is therefore a set of ordered pairs:

$$\text{Graph}(f) = \{(x, y) : y = 2x^3\}$$

For example, the points $(0, 0), (1, 2), (2, 16), (3, 54)$ belong to the graph. Now we assert that a function *is* its graph. Thus the function $f(x) = 2x^3$ is nothing but the set $\{(x, y) : y = 2x^3\} \subseteq \mathbb{R} \times \mathbb{R}$.

Examples A.2.17 You've already met more than just a few functions in your mathematical education up to date. The most obvious ones are functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, such as $f(x) = x^2, g(x,y) = \sin(x^3+y), h(x,y,z) = (xy, x\ln z)$, etc. Here are a few more that you might not yet have considered as functions:

(a) Define $\mathbb{Z} \xrightarrow{f} \mathcal{P}(\mathbb{Z})$ by: $f(n) = \{m : m \text{ divides } n\}$. Then f is a function which maps a number to a set. For example,
$$f(12) = \{\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12\} = f(-12)$$

(b) Let $\mathcal{C}^0(\mathbb{R},\mathbb{R}) = \{f : f \text{ is a continuous map from } \mathbb{R} \text{ to } \mathbb{R}\}$, and let $a \leq b \in \mathbb{R}$. Then $\int_a^b : \mathcal{C}^0(\mathbb{R},\mathbb{R}) \longrightarrow \mathbb{R}$ is a function which assigns to every continuous map its definite integral.

(c) Let $\mathcal{C}^1(\mathbb{R},\mathbb{R})$ be the set of all maps from $\mathbb{R}$ to $\mathbb{R}$ which have continuous first derivatives. Then the derivative operator is a map $D : \mathcal{C}^1(\mathbb{R},\mathbb{R}) \longrightarrow \mathcal{C}^0(\mathbb{R},\mathbb{R})$.

(d) **curl** is a map from the set of vector fields on $\mathbb{R}^3$ to itself. **div** is a map from the set of vector fields on $\mathbb{R}^3$ to the set of functions on $\mathbb{R}^3 \to \mathbb{R}$. **grad** is a map from the set of differentiable functions $\mathbb{R}^3 \to \mathbb{R}$ to the set of vector fields on $\mathbb{R}^3$.

(e) An $n \times m$ matrix A can be regarded as a map from $A : \mathbb{R}^m \longrightarrow \mathbb{R}^n$.

(f) Addition and multiplication are functions from $\mathbb{R}^2$ to $\mathbb{R}$. Addition can, in fact, be described by the 1×2–matrix $(1\ \ 1)$, for $(1\ \ 1)\binom{a}{b} = a + b$.

(g) If Ω is a universal set, then union and intersection can be regarded as functions from $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ to $\mathcal{P}(\Omega)$, which map the ordered pair (A, B) to $A \cup B$ and $A \cap B$ respectively.

(h) We can also regard the bigger version $\bigcup$ of union as a map, but this time we have $\bigcup : \mathcal{P}(\mathcal{P}(\Omega)) \longrightarrow \mathcal{P}(\Omega)$. It assigns to any family of subsets of Ω its union. (Note that a family of subsets of Ω is just a set of elements of $\mathcal{P}(\Omega)$, i.e. it is a subset of $\mathcal{P}(\Omega)$, and therefore an element of $\mathcal{P}(\mathcal{P}(\Omega))$.) The same goes for intersection.

□

For any set A, there is an important function on A called the *identity function*. It is denoted by id_A, and is defined by

$$\mathrm{id}_A : A \longrightarrow A \qquad \mathrm{id}_A(a) = a$$

Thus $\mathrm{id}_A = \{(a,a) : a \in A\}$.

Examples A.2.18 (a) The identity function on $\mathbb{R}$ is just the function $y = x$.

(b) The identity function on $\mathbb{R}^n$ is the identity matrix

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

□

Definition A.2.19 Let $f : A \to B$. If $A' \subseteq A$, we can define the *restriction* of f to A' as follows:

$f|A'$ is a map from A' to B, such that $(f|A')(a) = f(a)$ for all $a \in A'$

Definition A.2.20 Let $A \xrightarrow{f} B$ be a function.

(a) f is said to be *one-to-one* (or 1-1, or *injective*) if and only if the following condition holds:

If $f(a_1) = f(a_2)$, then $a_1 = a_2$.

(b) f is said to be *onto* (or *surjective*) if and only if

For every $b \in B$ there exists an $a \in A$ such that $f(a) = b$.

(c) f is said to be a *bijection* (or a *one-to-one correspondence*) if it is both an injection and a surjection.

Remarks A.2.21 A function $f : A \to B$ is injective if no two distinct members of a map to the same $b \in B$, i.e. if every $b \in B$ has *at most one* preimage.
f is surjective if and only if every b in B gets mapped onto by some $a \in A$, i.e. if every $b \in B$ has *at least one preimage*. In that case B is the range of f, i.e. $\text{ran}(f) = \text{codom}(f)$.
f is a bijection if and only if every $b \in B$ has *exactly one* preimage.

It should be clear that there is a bijection from a finite set A to another set B if and only if A and B have the same number of elements.

□

Examples A.2.22 (a) Let $f(x) = x^2$. We would generally regard f as a function with domain $\mathbb{R}$ and codomain $\mathbb{R}$. The range of f is $[0, +\infty)$, since f takes no negative values. f is not injective, because, for example $f(1) = f(-1)$. f is not surjective either, since -1 is not in the range of f.

(b) If we define $g(x) : [0, 1] \longrightarrow [0, 1]$ by $g(x) = x^2$, then we may regard g as the restriction of f to $[0, 1]$, i.e. $g = f|[0, 1]$. Now g is clearly a bijection.

(c) $x^3 : \mathbb{R} \longrightarrow \mathbb{R}$ is a bijection.

(d) Let $\mathbb{Q}^+$ denote the set of all non–negative rational numbers. The map $h : \mathbb{Z} \times \mathbb{N} \longrightarrow \mathbb{Q}^+$ defined by $h(n, m) = \frac{n}{m}$ is surjective, but not injective.

(e) If $A \subseteq B$, then the **inclusion** $f : A \to B$ defined by: $f(a) = a$ is an injection. It is a bijection if and only if $A = B$.

(f) Let A be an $n \times n$–matrix, regarded as a map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then A is injective if and only if $\det(A) \neq 0$.

□

Next, we discuss how functions can be combined:

Definition A.2.23 If $f : A \to B$ and $g : B \to C$, then $g \circ f$ is a function from A to C, defined by

$$(g \circ f)(a) = g(f(a))$$

Note that the composition $g \circ f$ does in one step what f and g do in two:

$$A \xrightarrow{f} B \xrightarrow{g} C \qquad a \overset{f}{\mapsto} f(a) \overset{g}{\mapsto} g(f(a))$$
$$A \xrightarrow{g\circ f} C \qquad a \overset{g\circ f}{\mapsto} g(f(a))$$

Also note that $g \circ f$ means:

Do f first, then g

i.e. the last shall be first.

An often used fact is that *composition is an associative operation* on functions, i.e.

$$h \circ (g \circ f) = (h \circ g) \circ f$$

By this equation we mean that: one side is defined if and only if the other side is defined, and in that case they are equal.
For if $A \xrightarrow{f} B, B \xrightarrow{g} C$, and $C \xrightarrow{h} D$, then $h \circ (g \circ f)$ is a function from A to D which works as follows: First do $g \circ f$, then do h. But to do $g \circ f$, you must first do f, then g. The combined result is

First do f, then g, and then h: $(h \circ (g \circ f))(a) = h(g(f(a)))$

Similarly, $(h \circ g) \circ f$ is a function from A to D which works as follows: First do f, then $h \circ g$. But to do $h \circ g$, you must first do g, then h. The combined result is therefore

First do f, then g, and then h: $((h \circ g) \circ f)(a) = h(g(f(a)))$

and thus $h \circ (g \circ f) = (h \circ g) \circ f$, as claimed.

Example A.2.24 Consider the following functions (note their domains and codomains):

$$\mathbb{R} \xrightarrow{f} \mathbb{R}^+ : x \mapsto x^2 + 1$$
$$\mathbb{R}^+ \xrightarrow{g} \mathbb{R}^+ : y \mapsto \sqrt{y}$$
$$\mathbb{R}^+ \xrightarrow{h} [-1, 1] : z \mapsto \sin(z)$$

Then

$$\mathbb{R} \xrightarrow{g\circ f} \mathbb{R}^+ : x \mapsto \sqrt{x^2 + 1}$$
$$\mathbb{R}^{+1} \xrightarrow{h\circ g} [-1, 1] : y \mapsto \sin(\sqrt{y})$$

and thus

$$\mathbb{R} \xrightarrow{h\circ(g\circ f)} [-1, 1] : x \mapsto \sin(\sqrt{x^2 + 1})$$
$$\mathbb{R} \xrightarrow{(h\circ)g\circ f} [-1, 1] : x \mapsto \sin(\sqrt{x^2 + 1})$$

□

Exercises A.2.25 (1) Let $f : \mathbb{N} \to \mathbb{N} : n \mapsto n^2$, and let $g : \mathbb{N} \to \mathbb{N} : n \to n + 2$. Calculate $(f \circ g)(5)$ and $g \circ f(5)$.
Write down formulas for $f \circ g$ and $g \circ f$.

(2) Suppose that $f(x) = x^2$ and $g(x) = x + 3$. Calculate $g \circ f(x)$ and $f \circ g(x)$. Note that $g \circ f \neq f \circ g$.

(3) If A is an $n \times m$–matrix, and B is an $m \times r$–matrix, then we can regard them as functions $\mathbb{R}^m \xrightarrow{A} \mathbb{R}^n$, $\mathbb{R}^r \xrightarrow{B} \mathbb{R}^m$. The composition $A \circ B$ is therefore a map $\mathbb{R}^r \to \mathbb{R}^n$. It is not hard to show that the composition is just the matrix product, i.e. that $A \circ B = AB$. Do so!

(4) Suppose that $g \circ f_1 = g \circ f_2$. Prove that if g is injective then we can "cancel" g to conclude $f_1 = f_2$. Give an example to show that left–cancellation may fail if g is not injective.

(5) Suppose that $g_1 \circ f = g_2 \circ f$. Prove that if f is surjective then we can "cancel" f to obtain $g_1 = g_2$. Show that right–cancellation may fail if f is not surjective.

□

Note that if $f : A \longrightarrow B$, then $f \circ \mathrm{id}_A = f$, and $\mathrm{id}_B \circ f = f$. Thus the identity function behaves like an identity element for the operation of composition.

The number 0 is an identity element for the operation of addition, because $x + 0 = x$.
The number 1 is an identity element for the operation of multiplication, because $x \cdot 1 = x$.

Next, we tackle the idea of *inverting* (or *reversing*) the effect of a function. Take the function $f(x) = 3x$. It transforms the number x into the number $3x$. To *undo* this transformation, you just multiply $3x$ by $\frac{1}{3}$. The function $g(x) = \frac{1}{3}x$ inverts the effect of f, in that

$$g \circ f(x) = x \qquad f \circ g(y) = y$$

Thus applying first f, and then g gets you back to the starting point x. The same holds true if you apply g first, and then f.

Can every function be inverted? No, as is easy to see: Consider the function $f(x) = x^2$. Then $f(2) = 4 = f(-2)$. Now if g is a function which reverses the effect of f, then we cannot decide whether $g(4) = 2$ or $g(4) = -2$. The problem arises because g is not 1-1.

Let's make the preceding discussion precise:

Definition A.2.26 Let $f : A \to B$. We say that f is *invertible* if and only if there is a function $g : B \to A$ such that

$$g(f(a)) = a \quad \text{for all } a \in A, \qquad f(g(b)) = b \quad \text{for all } b \in B \tag{$*$}$$

The function g, *if it exists*, is called the *inverse* of f, and denoted $g = f^{-1}$. Then $(*)$ amounts to saying

$$f^{-1} \circ f = \mathrm{id}_A \qquad \text{and} \qquad f \circ f^{-1} = \mathrm{id}_B$$

Note that if f^{-1} exists, then

$$f^{-1}(b) = a \quad \text{if and only if} \quad f(a) = b$$

Proposition A.2.27 *A function $f : A \longrightarrow B$ is invertible if and only if it is a bijection.*

Proof: Suppose that f is invertible, i.e. that f^{-1} exists. Then f^{-1} is a function from B to A. We first show that f is surjective: Let $b \in B$. Since the domain is B, $f^{-1}(b)$ must be defined, i.e. there must be some $a \in A$ such that $f^{-1}(b) = a$. But then $f(a) = b$. Hence every $b \in B$ has a preimage.
Next we show that f is injective. For suppose that $f(a_1) = f(a_2) = b$. Then $f^{-1}(b) = a_1$ and $f^{-1}(b) = a_2$. Since f^{-1} is a function, we must have $a_1 = a_2$ (check the definition of function), and hence f is injective.
This proves that if f is invertible, then f is a bijection.

Now we prove the converse. If f is a bijection, then it is onto B. Hence for every $b \in B$ there is some $a \in A$ such that $f(a) = b$. Moreover, since f is one-to-one, that a has to be unique. So we may define $f^{-1}(b)$ to be the unique a such that $f(a) = b$. This makes f^{-1} into a well–defined function $f^{-1} : B \to A$.

⊣

Examples A.2.28 (a) The function $f(x) = x^3$ is a bijection on the reals, and its inverse is $g(x) = \sqrt[3]{x}$.

(b) The function $f(x) = x^2$ does not have an inverse, since it is not a bijection. However, if we *restrict* f to the non–negative reals, then $f|\mathbb{R}^+$ is a bijection. Its inverse is the square root function.

(c) The function $f : \mathbb{R} \longrightarrow (0, +\infty)$ defined by $f(x) = e^x$ is bijective. Its inverse is the natural logarithm $\ln x$.

(d) The function $\sin x$ is neither injective, nor surjective; however, if we restrict $\sin x$ and regard it as a function $[-\frac{\pi}{2}, \frac{\pi}{2}] \longrightarrow [-1, 1]$, then it is a bijection, and its inverse is $\arcsin x$.

(e) If A is an $n \times n$–matrix, regarded as a function on $\mathbb{R}^n$, then A has an inverse function if and only if A has an inverse matrix. Since composition is just matrix multiplication, the *inverse function* of A is just the *inverse matrix* A^{-1}.

□

Remarks A.2.29 Note that, in general,

$$f^{-1}(x) \quad \neq \quad \frac{1}{f(x)}$$

e.g. $\sqrt[3]{x} \neq \frac{1}{x^3}$.

The number $x^{-1} = \frac{1}{x}$ is the inverse of x under the operation of *multiplication*, in that

$$x \cdot x^{-1} = 1 \qquad x^{-1} \cdot x = 1$$

noting that 1 is the identity for multiplication.
The function f^{-1} is the inverse of f under the operation of *composition*, in that

$$f \circ f^{-1} = \mathrm{id} \qquad f^{-1} \circ f = \mathrm{id}$$

noting that id is the identity for composition.
The same notation for inverse, i.e. $^{-1}$, refers to *different operations*, so there's no reason to believe that there is any relationship between them.

□

The notion of invertibility can be refined:

Definition A.2.30 Let $f : A \to B$ and $g : B \to A$.

(a) g is called a *left inverse* of f if $g \circ f = \mathrm{id}_A$.

(b) g is called a *right inverse* of f if $f \circ g = \mathrm{id}_B$.

Note that if f is invertible, then f^{-1} is both a left and a right inverse of f, and vice versa.

Exercises A.2.31 (1) Prove that a function f has a left inverse if and only if it is injective.

(2) Prove that a function f has a right inverse if and only if it is surjective.

(3) Prove that if a function f has a left inverse g *and* a right inverse h, then f is invertible, and $g = h$.

(4) Consider $f : \{a, b, c\} \to \{1, 2\}$ defined by $f(a) = f(b) = 1, f(c) = 2$. Find two distinct right inverses of f.

(5) Consider the inclusion $\iota : \mathbb{Z} \to \mathbb{Q}$. Construct two distinct left inverses of ι.

□

A.2.3 Functions Operating On Sets

We have already noted the confusion that may possibly arise by the two uses of the symbol $^{-1}$. We have but few symbols at our disposal, and many of them must therefore serve more than one function. Thus you must *always be aware of the context* in which a particular symbol is used.

You have to do this when using ordinary language: You *know* in what sense the newspaper headline

"School kids make great snacks at fund raiser"

is meant, even though the other sense offers greater amusement value.

I say this because we are about to add to the possible confusion. With every function $f : A \to B$ (not necessarily invertible), we can associate two new functions between the power sets of A and B

$$f[\cdot] : \mathcal{P}(A) \to \mathcal{P}(B) : A' \mapsto \{b \in B : \text{ There is } a' \in A' \text{ such that } f(a') = b\} \quad \text{where } A' \subseteq A$$
$$f^{-1}[\cdot] : \mathcal{P}(B) \to \mathcal{P}(A) : B' \mapsto \{a \in A : f(a) \in B'\} \quad \text{where } B' \subseteq B$$

Thus $f[\cdot]$ assigns to each subset A' of A a subset $f[A'] \subseteq B$. Similarly, $f^{-1}[\cdot]$ transforms each subset B' of B into a subset $f^{-1}[B'] \subseteq A$.
We will, for the moment, use square brackets to distinguish the various functions, but will drop this convention later. Which function is meant will be clear from context. We shall also call $f[A']$ the *direct image* of A' along f, and $f^{-1}[B']$ the *inverse image* of B' along f. Note that

$$f[A'] = \text{set of all images of } a \in A'$$

whereas

$$f^{-1}[B'] = \text{set of all preimages of } b \in B'$$

Remarks A.2.32 Sometimes the notation $f^{\rightarrow}$ is used for direct image, and $f^{\leftarrow}$ for inverse image.

□

Inverse images play a very important role in mathematics. It is therefore useful to remember the following:

$$a \in f^{-1}[B'] \quad \text{if and only if} \quad f(a) \in B'$$

Similarly,

$$b \in f[A'] \quad \text{if and only if there is } a' \in A' \text{ such that } f(a') = b$$

Examples A.2.33 (a) Suppose that $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$. Then

$$f[-1,2] = [0,4] \quad , f[\mathbb{Z}] = \{0,1,4,9,\dots\}, \quad f[\{4\}] = \{16\}$$

Also

$$f^{-1}[0,1] = [-1,1], \quad f^{-1}[\{4\}] = \{2,-2,\}, \quad f^{-1}[\{-4\}] = \emptyset$$

In each case, a *set* is transformed into a *set*.

(b) Suppose that $A = \{a_1, a_2, a_3\}, B = \{b_1, b_2, b_3\}$, and that $f : A \to B$ is defined by $f(a_1) = f(a_3) = b_1$, and $f(a_2) = b_3$. Then

$$f[\{a_1\}] = f[\{a_3\}] = f[\{a_1, a_3\}] = \{b_1\}, \quad f[\{a_2\}] = b_3, \quad f[A] = \{b_1, b_3\}, \quad f[\emptyset] = \emptyset$$

and

$$f^{-1}[\{b_3\}] = \{a_2\}, \quad f^{-1}[\{b_2\}] = f^{-1}[\emptyset] = \emptyset, \quad f^{-1}(B) = f^{-1}[\{b_1, b_3\}] = A$$

□

Exercises A.2.34 1. Let $f : A \to B$ be a function, and let $A' \subseteq A$, $B' \subseteq B$.

(a) Show that $A' \subseteq f^{-1}[f[A']]$

(b) Show that $B' \supseteq f[f^{-1}[B']]$

(c) Show that $A' = f^{-1}[f[A']]$ for every A' if and only if f is injective.

(d) Show that $B' = f[f^{-1}[B']]$ for every B' if and only if f is surjective.

[Hints: Reason along the following lines:
(b) If $b \in f[f^{-1}[B']]$ then $b = f(a)$ for some $a \in f^{-1}[B']$. But then $f(a) \in B'$, and so $b \in B'$.
(c) If $a \in f^{-1}[f[A']]$ then $f(a) \in f[A']$. Thus there is $a' \in A'$ such that $f(a) = f(a')$. But since f is injective, $a = a'$, and so $a \in A'$.]

2. Inverse images preserve the set operations: Let $f : A \to B$, and suppose that G, H are subsets of B. Then

(a) If $G \subseteq H$, then $f^{-1}[G] \subseteq f^{-1}[H]$;

(b) $f^{-1}[G \cap H] = f^{-1}[G] \cap f^{-1}[H]$;

(c) $f^{-1}[G \cup H] = f^{-1}[G] \cup f^{-1}[H]$;

(d) $f^{-1}[G - H] = f^{-1}[G] - f^{-1}[H]$;

3. Direct images are not quite so well behaved: Let $f : A \to B$, and suppose that $G, H \subseteq A$.
   (a) Suppose that $G \subseteq H$. Show that $f[G] \subseteq f[H]$;
   (b) Show that $f[G \cup H] = f[G] \cup f[H]$;
   (c) Show that $f[G \cap H] \subseteq f[G] \cap f[H]$;
   (d) Give an example to show that we may not have $f[G \cap H] = f[G] \cap f[H]$;
   (e) Show that $f[G] - f[H] \subseteq f[G - H] \subseteq f[G]$;
   (f) Give an example to show, in (e), that both $\subseteq$'s may fail to be $=$'s.

□

We end this section with some notation: Suppose that A, B are finite sets, and that A has n elements, and B m elements. How many functions are there from A to B?
For each $a \in A$ we have m choices for the value $f(a) \in B$. Thus there are m^n functions from A to B. For that reason

Definition A.2.35 Let A, B be sets. Then we define

$$B^A = \text{set of all functions from } A \text{ to } B$$

Some authors use ${}^A B$ instead of B^A.

□

Note that each function $f : A \to B$ is a subset of $A \times B$. Hence B^A is a set of subsets of $A \times B$, i.e. $B^A \in \mathcal{P}(\mathcal{P}(A \times B))$.

A.3 Countable and Uncountable Sets

In this section, we investigate the idea of the *size* or *cardinality* of a set. For finite sets, we can determine the size of a set by counting its elements. Thus for example, the set $\{a, b, c\}$ has cardinality 3 (it has 3 elements). We are going to extend this idea of counting to obtain the size to infinite sets, and we will show that infinity comes in many sizes.

Let's explore the idea of *counting*: For the moment, let $\mathbf{n} = \{1, 2, \ldots, n\}$ be the set of the first n natural numbers. To say that $A = \{a, b, c\}$ has 3 elements is equivalent to saying that there is a one-to-one correspondence between the sets A and $\mathbf{3}$. Indeed, this is the heart of the idea of counting: When we count the elements of A, we are setting up a bijection between A and $\mathbf{3}$. We go "a first, b second, c third". This is equivalent to a map $f : A \cong \mathbf{3}$ defined by $f(a) = 1, f(b) = 2, f(c) = 3$. Thus the idea of counting the elements of a finite set X involves finding a bijection between X and some $\mathbf{n}$. If there is a bijection from X to $\mathbf{n}$, then X has n elements.

It is obvious that two finite sets A and Δ have the same size if and only if there is a one-to-one correspondence $f : A \cong \Delta$. We don't even have to count A and Δ to know that they have the same number of elements. If $A = \{a, b, c, d\}$ and $\Delta = \{\alpha, \beta, \gamma, \delta\}$, then the existence of the bijection $f : A \cong \Delta$ given by

$$f(a) = \beta, f(b) = \delta, f(c) = \alpha, f(d) = \gamma$$

is sufficient to show that A and Δ have the same number of elements. It doesn't tell us that this number is 4.

Thus two sets have the same size if and only if there is a bijection between them; we can bypass the idea of number. This is important, because we cannot actually *count* infinite sets. But we can establish bijective correspondences between infinite sets. We shall adopt this idea as our basic idea of size.

Definition A.3.1 We define an equivalence relation $\approx$ between sets as follows: If A, B are sets, we say that $A \approx B$ if and only if there is a bijection from A to B. If $A \approx B$, we say that A and B have the same *cardinality*. We may also indicate this by saying $|A| = |B|$.

Note that having the same cardinality is an *equivalence relation* between sets, i.e. that

(i) $|A| = |A|$ (Reflexivity)

(ii) If $|A| = |B|$, then $|B| = |A|$ (Symmetry)

(iii) If $|A| = |B|$ and $|B| = |C|$, then $|A| = |C|$ (Transitivity)

Exercise A.3.2 Prove this assertion. (Note that the assertion is *not obvious*: When we say that $|A| = |B|$, we are not actually claiming that there are two equal numbers. What we *are* saying is that there is a bijection from A to B. To prove (i), for example, you have to find a bijection from A to A.)

□

Examples A.3.3 (a) Two finite sets have the same cardinality if and only if they have the same number of elements.

(b) For finite sets, if A is a *proper subset* of B, then $|A| < |B|$. This breaks down completely for infinite sets. Consider, for example, the sets $\mathbb{N}$ and $\mathbb{Z}$. It is certainly true that $\mathbb{N} \subset \mathbb{Z}$. However, the map $\mathbb{N} \xrightarrow{f} \mathbb{Z}$ defined by

$$f(n) = \begin{cases} \dfrac{n}{2} & \text{if } n \text{ is even} \\ -\dfrac{n-1}{2} & \text{if } n \text{ is odd} \end{cases}$$

is a bijection: $f(1) = 0, f(2) = 1, f(3) = -1, f(4) = 2, f(5) = -2, f(6) = 3 \ldots$. (Note that we are zig–zagging from the positive integers to the negative integers.) Thus $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality, even though $\mathbb{N}$ seems to contain fewer elements than $\mathbb{Z}$.

(c) We also have $|\mathbb{Q}| = |\mathbb{N}|$. This can be seen as follows. Put the set of strictly positive rational numbers $\mathbb{Q}^+$ in an array

$$\begin{array}{ccccccc} 1/1 & 2/1 & 3/1 & 4/1 & 5/1 & \ldots \\ 1/2 & 2/2 & 3/2 & 4/2 & 5/2 & \ldots \\ 1/3 & 2/3 & 3/3 & 4/3 & 5/3 & \ldots \\ 1/4 & 2/4 & 3/4 & 4/4 & 5/4 & \ldots \\ 1/5 & 2/5 & 3/5 & 4/5 & 5/5 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

We can then trace a zig–zag path that moves through all the rational numbers as follows. Start at the top line and move diagonally down to the left until you reach the leftmost line. Repeat. We thus obtain a sequence

$$\frac{1}{1},\ \frac{2}{1},\ \frac{1}{2},\ \frac{3}{1},\ \frac{2}{2},\ \frac{1}{3},\ \frac{4}{1},\ \frac{3}{2},\ \frac{2}{3},\ \frac{1}{4},\ \frac{5}{1}\ldots$$

All of the strictly positive rational numbers occur in this sequence, and they all occur infinitely many times. For example, $\frac{1}{1}, \frac{2}{2}, \frac{3}{3}\ldots$ lie along the diagonal, and they are all equal. To obtain a bijection from $\mathbb{N}$ to $\mathbb{Q}^+$, we follow the above sequence of rationals, but we omit any number that has already occurred to ensure that the function is one-to-one, i.e. we *prune* away the repeated values. We therefore define the function $\mathbb{N} \xrightarrow{f} \mathbb{Q}^+$ by

$$f(1) = \frac{1}{1},\ f(2) = \frac{2}{1},\ f(3) = \frac{1}{2},\ f(4) = \frac{3}{1},\ f(5) = \frac{1}{3},\ f(6) = \frac{4}{1}, \ldots$$

Note that $f(5) \neq \frac{2}{2}$, which is after $f(4) = \frac{3}{1}$ in the sequence, because $\frac{2}{2} = \frac{1}{1}$ has already occurred as $f(1)$. Then f is a bijection from $\mathbb{N}$ to $\mathbb{Q}^+$. Now even though we haven't found a *formula* for f, it is nevertheless a perfectly good function, and all its values can be calculated. Can you see that $f(16) = \frac{2}{5}$?

In the same way, we can set up a bijection g from $\mathbb{N}$ to the negative rationals. Just put $g(n) = -f(n)$. Finally, we can define a bijection $h : \mathbb{N} \longrightarrow \mathbb{Q}$ using f, g and another zig–zag: We define

$$h(1) = 0,\ h(2) = f(1),\ h(3) = g(1),\ h(4) = f(2),$$
$$h(5) = g(2),\ h(6) = f(3),\ h(7) = g(3), \ldots$$

Again, we have no formula for h, but it is certainly a well–defined function, and all its values can be calculated. Check that $h(23) = -\frac{1}{5}$.

□

Definition A.3.4 A set A is said to be *countable* if there is a surjection from $\mathbb{N}$ onto A.

Remarks A.3.5 (a) Basically a set A is countable if its elements can be indexed by the natural numbers, i.e. if it *can* be written as $A = \{a_n : n \in \mathbb{N}\}$. For if A is countable and not finite, then there is a bijection $\mathbb{N} \xrightarrow{f} A$, and we can take $a_n = f(n)$. Conversely, if $A = \{a_n : n \in \mathbb{N}\}$ is infinite, we can define a bijection from $\mathbb{N}$ to A by letting $f(n) = a_n$ (although here some *pruning* is necessary if the a_n aren't all distinct; see Example A.3.3(c)).

(b) A set A is countable if and only if it is either finite or can be put into a one-to-one correspondence with the natural numbers, i.e. if $|A| = \mathbf{n}$ for some $n \in \mathbb{N}$, or $|A| = |\mathbb{N}|$.

(c) In Example A.3.3, we proved that the sets $\mathbb{Z}$ and $\mathbb{Q}$ are countable sets.

(d) The "zig–zag" technique, used above to prove that the rational numbers are countable, is often very useful.

□

Exercise A.3.6 Prove that the union of countably many countable sets is countable (i.e. prove that if A_n ($n \in \mathbb{N}$) are countable sets, then the set $\bigcup_{n\in\mathbb{N}} A_n$ is countable as well.) [Hint:Zig–zag!]

□

So all the infinite sets we've seen so far are countable (and the finite ones also, of course). A very natural question that might occur to you is the following: Are all infinite sets countable? The answer is **"No!"**

Example A.3.7 We show that the unit interval $I = [0, 1]$ is *uncountable*, i.e. that we cannot find an enumeration

$$I = \{x_n : n \in \mathbb{N}\}$$

The proof is by *contradiction*: Suppose that we *can* find such an enumeration $I = \{x_1, x_2, x_3, x_4, \dots\}$, i.e that every real number in $[0, 1]$ is equal to x_n for some n. Now every number x_n has a decimal expansion of the form

$$x_n = 0.x_{n1}x_{n2}x_{n3}x_{n4}x_{n5}\dots$$

where x_{nm} is the m^{th} number in the decimal expansion of x_n. Of course some real numbers have two distinct decimal expansions, a terminating one and a non–terminating one. For example, $1.0000\cdots = 0.9999\dots$[2] We will choose the non–terminating decimal expansions for our x_n.

We now create a new real number x from the x_n by a process called *diagonalization*. We choose $a_n \in \{1, 2, \dots, 9\}$ such that the following hold:

$$a_1 \neq x_{11}, a_2 \neq x_{22}, a_3 \neq x_{33}, \dots, a_n \neq x_{nn}, \dots$$

To avoid a situation where we obtain a number x with a terminating decimal expansion, we haven't permitted $a_n = 0$; this is just a technicality. We can now define x: Put

$$x = 0.a_1a_2a_3a_4\dots$$

Here comes the heart of the argument: Clearly $x \in I = [0, 1]$. Now if I can be written as a list $\{x_1, x_2, x_3, \dots\}$, then there must be some n such that $x = x_n$. But the first decimal place of x differs from the first decimal place of x, since $a_1 \neq x_{11}$; hence $x \neq x_1$. Similarly, the second decimal place of x differs from the second decimal place of x_2, since $a_2 \neq x_{22}$; hence $x \neq x_2$. We can continue in this way to show that $x \neq x_n$ for any $n \in \mathbb{N}$, i.e. x is not on the list $\{x_1, x_2, x_3, \dots\}$.

This proves the result! Given any list $x_1, x_2, x_3, \dots$ of real numbers in $[0, 1]$, we now have a technique for producing a new real number x that is not on the list. It thus follows that no such list can contain all the real numbers in $[0, 1]$, i.e. there is no bijection from $\mathbb{N}$ to $[0, 1]$.

□

Hence there are uncountable sets. Clearly $\mathbb{R}$ is also uncountable, because otherwise we could find an enumeration $\{r_1, r_2, r_3, \dots\}$ of $\mathbb{R}$. By omitting any reals which are not in $[0, 1]$, we could prune this into an enumeration of $[0, 1]$.

[2] An easy way to see this is to note that $1 = 3 \times \frac{1}{3} = 3(0.333\dots) = 0.999\dots$.

Exercise A.3.8 Show that if A is any set, then $|A| \neq |\mathcal{P}(A)|$. Conclude that $\mathcal{P}(\mathbb{N})$ is uncountable. (Actually, it can be proved that $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$.)
[Hint: Suppose that $f : A \to \mathcal{P}(A)$ and consider the set $B := \{a \in A : a \notin f(a)\}$. By contradiction, show that $B \notin \mathrm{im} f$.]

□

Consider that

1. $\mathbb{Q}$ satisfies all the axioms that $\mathbb{R}$ does, except for the Completeness Axiom;

2. $\mathbb{Q}$ is countable, but $\mathbb{R}$ is uncountable.

This juxtaposition leads one to suspect that it is the Completeness Axiom which is responsible for the uncountability of $\mathbb{R}$. This is indeed the case, as you will see by proving the following proposition:

Proposition A.3.9 *Let I be a non–empty bounded interval of real numbers, and let $\{a_n : n \in \mathbb{N}\} \subseteq I$. Then there is a $p \in I$ such that $p \neq a_n$ for all n.*
In particular, $I \neq \{a_n : n \in \mathbb{N}\}$ for any sequence a_n. Hence I is uncountable.

Exercise A.3.10 We prove Propn. A.3.9:

(a) Explain why there is a *closed* interval $I_1 \subseteq I$ such that $a_1 \notin I_1$. [Hint: Divide I into three closed subintervals of equal length.]

(b) Explain why there is a closed subinterval $I_2 \subseteq I_1$ such that $a_2 \notin I_2$. Exlain why also $a_1 \notin I_2$.

(c) Now assume that we have found a closed subinterval I_n such that $a_1, \ldots, a_n \notin I_n$. Explain why there is a closed subinterval $I_{n+1} \subseteq I_n$ such that $a_{n+1} \notin I_{n+1}$. Explain also why we now have $a_1, a_2, \ldots, a_{n+1} \notin I_{n+1}$.

(d) We now have constructed a sequence of closed intervals

$$I_1 \supseteq I_2 \supseteq I_3 \cdots \supseteq I_n \supseteq \ldots$$

Explain why $\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset$.
[Hint: This uses the Completeness Axiom: Let l_n be the left endpoint of I_n. Show that $\sup\{l_n : n \in \mathbb{N}\}$ exists, and that $\sup\{l_n : n \in \mathbb{N}\} \in \bigcap_{n\in\mathbb{N}} I_n$.]

(e) Let $p \in \bigcap_{n\in\mathbb{N}} I_n$. Explain why $p \in I$, and why $p \neq a_n$ for any $n \in \mathbb{N}$.

□